




3 1761 10374366 2



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743662>

12-001



Government
Publications

9

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

June 2007

•

Volume 33

•

Number 1



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website at www.statcan.ca.

National inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Depository Services Program inquiries	1-800-700-1033
Fax line for Depository Services Program	1-800-889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Accessing and ordering information

This product, catalogue no. 12-001-XIE, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select Publications.

This product, catalogue no. 12-001-XPB, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered by

- Phone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.ca
- Mail Statistics Canada
Finance Division
R.H. Coats Bldg., 6th Floor
100 Tunney's Pasture Driveway
Ottawa (Ontario) K1A 0T6
- In person from authorised agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under About us > Providing services to Canadians.



Survey Methodology

A journal
published by
Statistics Canada

June 2007 • Volume 33 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2007

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:

Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

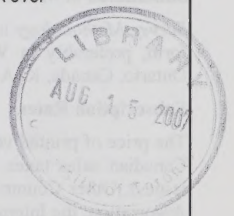
June 2007

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Past Chairmen G.J. Brackstone
R. Platek

Members J. Gambino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh

Associate Editors

D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidioglou, *Statistics Canada*
D. Judkins, *Westat Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
G. Nathan, *Hebrew University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*

J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
R. Valliant, *JPSM, University of Michigan*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 33, Number 1, June 2007

Contents

In This Issue.....	1
 Regular Papers	
Chris Skinner and Marcel de Toledo Vieira Variance estimation in the analysis of clustered longitudinal survey data	3
Milorad S. Kovačević and Georgia Roberts Modelling durations of multiple spells from longitudinal survey data	13
Michael R. Elliott Bayesian weight trimming for generalized linear regression models	23
F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson and M. Giovanna Ranalli Semiparametric model-assisted estimation for natural resource surveys	35
Marc Tanguay and Pierre Lavallée <i>Ex post</i> weighting of price data to estimate depreciation rates	45
David G. Steel and Robert G. Clark Person-level and household-level regression estimation in household surveys	51
Hiroshi Saigo Mean - Adjusted bootstrap for two - Phase sampling	61
Nicholas Tibor Longford On standard errors of model-based small-area estimators	69
Jun Shao Handling survey nonresponse in cluster sampling	81
Neeraj Tiwari, Arun Kumar Nigam and Ila Pant On an optimal controlled nearest proportional to size sampling scheme	87

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

∞

In This Issue

This issue of *Survey Methodology* includes papers covering a variety of methodological subjects such as modeling and estimation, weighting and variance estimation, non-response and sampling.

In the first paper of the issue, Skinner and Vieira investigate the effect of clustered sampling on variance estimation in longitudinal surveys. They present theoretical arguments and empirical evidence of the effects of ignoring clustering in longitudinal analyses, and find that these effects tend to be larger than for corresponding cross-sectional analyses. They also compare traditional survey sampling based methods to account for clustering in variance estimation to a multi-level modeling approach.

Kovačević and Roberts compare three models for analyzing multiple spells arising from data collected through longitudinal surveys with complex survey designs, which can involve stratification and clustering. These models are variations of the Cox proportional hazards model along the same lines as those proposed in the literature by Lin and Wei (1989), Binder (1992) and Lin (2000). These three models are compared using data from Statistics Canada's Survey on Labor and Income Dynamics (SLID). This paper gives new insight into fitting Cox models to survey data containing multiple spells per individual, a situation that arises quite frequently. The paper also illustrates some of the challenges in fitting Cox models to survey data.

Elliott, in his paper, presents a method for balancing elevated variance due to extreme weights with potential bias using a Bayesian weight trimming method in generalized linear models. This is accomplished by using a stratified hierarchical Bayesian model in which strata are determined by the probabilities of inclusion or survey weights. He illustrates and evaluates the approach using simulations based on linear and logistic regression models, and an application using data from the Partners for Child Passenger Safety dataset.

The paper by Breidt, Opsomer, Johnson and Ranalli explores the use of semiparametric methods for the estimation of population means. In semiparametric estimation, some variables are assumed to be linearly related to the variable of interest while the other variables may have a complicated, unspecified relation to the variable of interest. The authors study theoretically the properties under the sampling design of the resulting estimators. In particular, they show the design-consistency and the asymptotic normality of their estimator. Their method is then applied to data from a survey of lakes in the northeastern United States.

Tanguay and Lavallée address the problem of estimating the depreciation of assets based on a database of price ratios. In their paper, the issue is that the ratios do not come from a random sample from the population of ratios. The authors argue that the distribution of ratios should converge to a Uniform distribution and propose a weighting scheme that will make the weighted empirical distribution function approximately uniform. The proposed method is illustrated by an example using data on the depreciation of automobiles.

Steel and Clark present an empirical and theoretical comparison of person-level generalized regression survey weights and integrated household-level weights in the case of a simple random sample of households in which all household members selected. They conclude that there is little or no loss in efficiency associated with integrated weighting.

Saigo, in his paper, proposes a bootstrap variance estimation procedure for two-phase designs with high sampling fractions. The method uses common bootstrap techniques, but adjusts the values of the auxiliary variables for units that are selected in the first phase sample only. The proposed technique is illustrated using several commonly used estimators such as the ratio estimator, and estimators of the distribution function and quantiles. Results from a simulation study comparing the proposed method to several others are presented.

In the paper by Longford the problem of estimating the MSE of small area estimates is investigated. A composite estimator of the MSE of small area means is obtained by combining a model-based variance estimator and a naïve estimator of the MSE. The coefficient that combines the two estimators minimizes the expected MSE of the resulting composite estimator of the MSE. The proposed estimator is compared with existing estimators through several simulation studies.

Shao considers the problem of imputing for missing values when the nonresponse is nonignorable. In the situation where the nonresponse depends on a cluster level random effect, he shows that the commonly used mean imputed estimator is biased unless the mean of the cluster is used. For variance estimation, a jackknife variance estimation procedure for the proposed estimator is provided. The proposed estimator is compared with the mean imputed estimator by means of a simulation study.

In the final paper of this issue, Tiwari, Nigam and Pant make use of the idea of nearest proportional to size sampling designs to obtain optimal controlled sample designs where non-preferred samples have zero selection probabilities. The optimal controlled sampled designs are obtained by combining an initial inclusion probability proportional to size design and quadratic programming techniques to ensure that non-preferred samples have a zero selection probability. Their method is illustrated using several examples.

Harold Mantel, Deputy Editor

Variance estimation in the analysis of clustered longitudinal survey data

Chris Skinner and Marcel de Toledo Vieira¹

Abstract

We investigate the impact of cluster sampling on standard errors in the analysis of longitudinal survey data. We consider a widely used class of regression models for longitudinal data and a standard class of point estimators of a generalized least squares type. We argue theoretically that the impact of ignoring clustering in standard error estimation will tend to increase with the number of waves in the analysis, under some patterns of clustering which are realistic for many social surveys. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses. We illustrate this theoretical argument with empirical evidence from a regression analysis of longitudinal data on gender role attitudes from the British Household Panel Survey. We also compare two approaches to variance estimation in the analysis of longitudinal survey data: a survey sampling approach based upon linearization and a multilevel modelling approach. We conclude that the impact of clustering can be seriously underestimated if it is simply handled by including an additive random effect to represent the clustering in a multilevel model.

Key Words: Clustering; Design effect; Misspecification effect; Multilevel model.

1. Introduction

It is well known that it is important to take account of sample clustering when estimating standard errors in the analysis of survey data. Otherwise, standard error estimators can be severely biased. In this paper we investigate the impact of clustering in the regression analysis of longitudinal survey data and compare it with the impact on corresponding cross-sectional analyses. Kish and Frankel (1974) presented empirical work which suggested that the impact of complex designs on variances decrease for more complex analytical statistics and so one might conjecture that the impact on longitudinal analyses might also be reduced. We shall argue that, in fact, the impact of clustering on longitudinal analyses can tend to be greater, at least for a number of common types of analysis and for some common practical settings. An intuitive explanation is that some common forms of longitudinal analysis of individual survey data 'pool' data over time and enable much temporal 'random' variation in individual responses to be 'extracted' in the estimation of regression coefficients. In contrast, it may only be possible to extract much less variation in the effects of clustering since such clustering, representing geography for example, often tends to generate more stable effects than repeated measurements of individual behaviour. As a consequence the relative importance of clustering in standard errors can increase the more waves of data are included in the analysis.

In addition to considering the impact of clustering on variance estimation, we shall also consider the question of how to undertake the variance estimation itself. It is natural for many analysts to represent clustering via multilevel

models (Goldstein 2003, Chapter 9; Renard and Molenberghs 2002) and we shall consider how variance estimation methods based upon such models compare with survey sampling variance estimation procedures in the case of cluster sampling.

There is a well established literature on methods for taking account of complex sampling schemes in the regression analysis of survey data. See *e.g.*, Kish and Frankel (1974), Fuller (1975), Binder (1983), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003). We restrict attention here to 'aggregate' regression analyses (Skinner *et al.* 1989), where regression coefficients at the 'population level' are the parameters of interest, where suitable estimates of these coefficients may be obtained by adapting standard model-based procedures using survey weights and where the variances of these estimated regression coefficients may be estimated by linearization methods (Kish and Frankel 1974; Fuller 1975). In this paper, we extend this work to the case when longitudinal survey observations are obtained, based upon an initial sample drawn according to a complex sampling scheme, focussing again on the case of a clustered design. We consider a standard class of linear regression models for such longitudinal data, as considered in the biostatistical literature (*e.g.*, Diggle, Heagerty, Liang and Zeger 2002), the multilevel modelling literature (*e.g.*, Goldstein 2003) and the econometric literature (*e.g.*, Baltagi 2001). We consider an established class of point estimators of a generalized least squares type, modified by survey weighting. For some applications of such methods to survey data, see Lavange, Koch and Schwartz (2001); Lavange, Stearns, Lafata, Koch and Shah (1996).

1. Chris Skinner, University of Southampton, United Kingdom; Marcel de Toledo Vieira, Universidade Federal de Juiz de Fora, Brazil.

The impact of a complex sampling scheme on variance estimation will be measured by the ‘misspecification effect’, denoted meff (Skinner 1989a), which is the variance of the point estimator of interest under the actual sampling scheme divided by the expectation of a specified variance estimator. This is a measure of the relative bias of the specified variance estimator. If it is unbiased then the meff will be one. If the actual sampling scheme involves clustering but the specified variance estimator is ‘misspecified’ by ignoring the clustering, then the expectation of the variance estimator will usually be less than the actual variance and the meff will be greater than one. This concept is closely related to that of the ‘design effect’ or deff of Kish (1965), defined as the variance of the point estimator under the given design divided by its variance under simple random sampling with the same sample size, a concept more relevant to the choice of design than to the choice of standard error estimator.

We shall illustrate our theoretical arguments with analyses of data from the British Household Panel Survey (BHPS) on attitudes to gender roles, where the units of primary analytic interest are individual women and the clusters consist of postcode sectors, used as primary sampling units in the selection of the first wave sample from an address register.

The framework, including the models and estimation methods, is described in Section 2. The theoretical properties of the variance estimation methods are considered in Section 3. Section 4 illustrates these properties numerically, using an analysis of BHPS data. Some concluding remarks are provided in Section 5.

2. Regression model, data and inference procedures

Consider a finite population $U = \{1, \dots, N\}$ of N units, assumed fixed across a series of occasions $t = 1, \dots, T$. We shall refer to the units as individuals, although our discussion is applicable more generally. Let y_{it} denote the value of an outcome variable for individual $i \in U$ at occasion t and let $y_i = (y_{i1}, \dots, y_{iT})'$ be the vector of repeated measurements. Let x_{it} denote a corresponding $1 \times q$ vector of values of covariates for individual i at occasion t and let $x_i = (x_{i1}, \dots, x_{iT})'$. We assume that the following linear model holds for the expectation of y_i conditional on (x_1, \dots, x_N) :

$$E(y_i) = x_i\beta, \quad (1)$$

where β is a $q \times 1$ vector of regression coefficients and the expectation is with respect to the model. We suppose that β is the target for inference, that is the regression coefficients are the parameters of primary interest to the analyst.

Although we shall consider further features of this model, such as the covariance matrix of y_i , these will be assumed to be of secondary interest to the analyst.

The data available to make inference about β are from a longitudinal survey in which values of y_{it} and x_{it} are observed at each occasion (wave) $t = 1, \dots, T$ for individuals i in a sample, s , drawn from U at wave 1 using a specified sampling scheme. For simplicity, we assume no non-response here, but return to this possibility in Section 4.

In order to formulate a point estimator of β , we extend the specification of (1) to the following ‘working’ model:

$$y_{it} = x_{it}\beta + u_i + v_{it}, \quad (2)$$

where u_i and v_{it} are independent random effects with zero means and variances $\sigma_u^2 = \rho\sigma^2$ and $\sigma_v^2 = (1 - \rho)\sigma^2$ respectively, conditional on (x_1, \dots, x_N) . This model may be called a uniform correlation model (Diggle *et al.* 2002, page 55) or a two-level model (Goldstein 2003). The parameter ρ is the intra-individual correlation.

The basic point estimator of β we consider is

$$\hat{\beta} = \left(\sum_{i \in s} w_i x_i' V^{-1} x_i \right)^{-1} \sum_{i \in s} w_i x_i' V^{-1} y_i, \quad (3)$$

where w_i is a survey weight and V is a $T \times T$ estimated covariance matrix of y_i under the working model (2), *i.e.*, it has diagonal elements $\hat{\sigma}^2$ and off-diagonal elements $\hat{\rho}\hat{\sigma}^2$, where $(\hat{\rho}, \hat{\sigma}^2)$ is an estimator of (ρ, σ^2) . (Note that in fact $\hat{\sigma}^2$ cancels out in (3) and hence σ^2 does not need to be estimated for $\hat{\beta}$). In the absence of the weight terms and survey considerations, the form of $\hat{\beta}$ is motivated by the generalized estimating equations (GEE) approach of Liang and Zeger (1986). The idea here is that $\hat{\beta}$, as a generalized least squares estimator of β , would be fully efficient if the working model (2) held. However, $\hat{\beta}$ remains consistent under (1) and may still be expected to combine within- and between-individual information in a reasonably efficient way even if the working model for the error structure does not hold exactly.

The survey weights are included in (3) following the pseudo-likelihood approach (Skinner 1989b) to ensure that $\hat{\beta}$ is approximately unbiased for β with respect to the model and the design, provided (1) holds.

There are a number of alternative ways of estimating ρ . In a non-survey setting, Liang and Zeger (1986) provide an iterative approach which alternates between estimates of β and ρ . Shah, Barnwell and Bieler (1997) describe how survey weights may be incorporated into this approach and implement this method in the REGRESS procedure of the software SUDAAN. By default, SUDAAN implements only one step of this iterative method and, in the non-survey setting, Lipsitz, Fitzmaurice, Orav and Laird (1994) conclude there is little to be lost by using only a single step.

For the working model in (2), the approach of Liang and Zeger (1986) to the estimation of β and ρ is virtually identical to the iterative generalized least squares (IGLS) estimation approach of Goldstein (1986). Both methods iterate between estimates of β and ρ and both use GLS to estimate β given the current estimate of ρ . The only slight difference is in the method used to estimate ρ . Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) show how to incorporate survey weights into the IGLS approach and their method may be expected to lead to very similar estimates of ρ to those in the SUDAAN REGRESS procedure. For the purposes of this paper, the precise form of $\hat{\rho}$ will not be critical and we may view $\hat{\beta}$ as either a weighted GEE or a weighted IGLS estimator.

We now turn to the estimation of the covariance matrix of $\hat{\beta}$ under the complex sampling scheme. We shall generally assume that a stratified multistage sampling scheme has been employed. We consider two main approaches to variance estimation.

Our first approach is the classical method of linearization (Skinner 1989b, page 78). The estimator of covariance matrix of $\hat{\beta}$ is

$$\begin{aligned} v(\hat{\beta}) = & \left[\sum_{i \in s} w_i x_i' V^{-1} x_i \right]^{-1} \\ & \times \left[\sum_h n_h / (n_h - 1) \sum_a (z_{ha} - \bar{z}_h)(z_{ha} - \bar{z}_h)' \right] \\ & \times \left[\sum_{i \in s} w_i x_i' V^{-1} x_i \right]^{-1} \end{aligned} \tag{4}$$

where h denotes stratum, a denotes primary sampling unit (PSU), n_h is the number of PSUs in stratum h , $z_{ha} = \sum_i w_i x_i' V^{-1} e_i$, $\bar{z}_h = \sum_a z_{ha} / n_h$ and $e_i = y_i - x_i \hat{\beta}$. Similar estimators are considered by Shah *et al.* (1997, pages 8-9) and Lavange *et al.* (2001). If the weights, the sampling scheme and the difference between $n/(n-1)$ and 1 are ignored, this estimator reduces to the 'robust' variance estimator presented by Liang and Zeger (1986).

Our second approach is more directly model-based. The model is first extended to represent the complex population underlying the sampling scheme and inference then takes place with respect to the extended model. We consider only the case of two-stage sampling from a clustered population, where the two-level model in (2) is extended to the three-level model (Goldstein 2003):

$$y_{ait} = x_{ait} \beta + \eta_a + u_{ait} + v_{ait}. \tag{5}$$

The additional subscript a denotes cluster and the additional random term η_a with variance σ_η^2 represents the cluster effect (assumed independent of u_{ait} and v_{ait}). We let σ_u^2 and σ_v^2 denote the variances of u_{ait} and v_{ait} respectively. Inference then takes place using IGLS, which may be

weighted to avoid selection bias. This approach generates an estimated covariance matrix of the estimator of β directly. It should be noted, however that the estimator of β derived using weighted IGLS under model (5) may differ slightly from the estimator in (3) (although, for given estimates of the three variance components in (5), it will be the same as a weighted GEE estimator with a working covariance matrix based on this three-level model). Nevertheless, from our experience of social survey applications, such as in Section 4, and from theory (Scott and Holt 1982) the difference between these alternative point estimators will often be negligible.

Two broad approaches to deriving variance estimators from (5) are available. First, ignoring survey weights, the standard IGLS method (Goldstein 1986) may be employed, assuming that each random effect follows a normal distribution. Second, to avoid the assumption of normal homoscedastic random effects, a 'robust' variance estimation method (Goldstein 2003, page 80) may be employed. This approach is extended to handle survey weights in Pfeffermann *et al.* (1998). Leaving aside stratification, their variance estimator is identical to the linearization estimator in (4) for a given value of $\hat{\rho}$.

3. Properties of variance estimators

In this section we consider the properties of the estimators of the covariance matrix of $\hat{\beta}$ described in the previous section. We focus first on the linearization estimator $v(\hat{\beta})$ in (4).

The consistency of $v(\hat{\beta})$ for the covariance matrix of $\hat{\beta}$ follows established arguments in a suitable asymptotic framework (e.g., Fuller 1975; Binder 1983). The one non-standard feature is the presence of V^{-1} in $\hat{\beta}$ and $v(\hat{\beta})$ and the dependence of V on $\hat{\rho}$. In fact, in large samples the covariance matrix of $\hat{\beta}$ depends on $\hat{\rho}$ only via its limiting value ρ^* (in a given asymptotic framework). To see this, write $\hat{\beta} - \beta = (\sum_s u_i)^{-1} \sum_s \tilde{z}_i$, where $u_i = w_i x_i' V^{-1} x_i$, $\tilde{z}_i = w_i x_i' V^{-1} \tilde{e}_i$ and $\tilde{e}_i = y_i - x_i \beta$. Note that, under weak regularity conditions (Fuller and Battese 1973, Corollary 3), the asymptotic distribution of $\hat{\beta} - \beta$ is the same as that of $\beta^* - \beta = (\sum_s u_i^*)^{-1} \sum_s z_i^*$, where $u_i^* = w_i x_i' V^{*-1} x_i$, $z_i^* = w_i x_i' V^{*-1} \tilde{e}_i$ and V^* takes the same form as V with $\hat{\rho}$ replaced by $\rho^* = p \lim(\hat{\rho})$, the probability limit of $\hat{\rho}$ in the asymptotic framework. Writing $\bar{z}^* = \sum_s z_i^* / n$ and $\bar{U} = p \lim(\sum_s u_i^* / n)$, we may thus approximate the covariance matrix of $\hat{\beta}$ asymptotically by $\text{var}(\hat{\beta}) \approx \bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$. If the working model (2) holds then $\rho^* = \rho$ and this covariance matrix will be the same for any consistent method of estimating ρ . Even if the working model does not hold, $v(\hat{\beta})$ will be consistent for $\bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$ within the kinds of asymptotic frameworks considered by

Fuller (1975) and Binder (1983) and under the kinds of regularity conditions they and Fuller and Battese (1973) set out.

We next explore the impact on the linearization method of ignoring a complex sampling design. We denote by $v_0(\hat{\beta})$ the linearization estimator obtained from expression (4) by ignoring the design, *i.e.*, by assuming only a single stratum with PSUs identical to individuals so that $n_h = n$ is the overall sample size and z_{ha} is replaced by $z_i = w_i x_i' V^{-1} e_i$. We shall be concerned with the bias of $v_0(\hat{\beta})$ when in fact the design is complex. Let $\hat{\beta}_k$ denote the k^{th} element of $\hat{\beta}$ and let $v_0(\hat{\beta}_k)$ denote the k^{th} element of $v_0(\hat{\beta})$. Then, following Skinner (1989a, page 24), we shall measure the relative bias of the 'incorrectly specified' variance estimator $v_0(\hat{\beta}_k)$ as an estimator of $\text{var}(\hat{\beta}_k)$ by the *misspecification effect*, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \text{var}(\hat{\beta}_k) / E[v_0(\hat{\beta}_k)]$. Since $v(\hat{\beta}_k)$ is a consistent estimator of $\text{var}(\hat{\beta}_k)$, $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be estimated by $v(\hat{\beta}_k) / v_0(\hat{\beta}_k)$ and is closely related to the idea of design effect.

To investigate the nature of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$, we first write:

$$v_0(\hat{\beta}) = (\sum_s u_i)^{-1} [n/(n-1)] \times [\sum_s (z_i - \bar{z})(z_i - \bar{z})'] (\sum_s u_i)^{-1}, \quad (6)$$

where $\bar{z} = \sum_s z_i / n$. Then, as an asymptotic approximation, we have $E[v_0(\hat{\beta})] \approx \bar{U}^{-1} [n^{-1} S_z^*] \bar{U}^{-1}$, where S_z^* is the probability limit of the finite population covariance matrix of z_i^* . Using the fact that the numerator of $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$ may be approximated by $\bar{U}^{-1} \text{var}(\bar{z}^*) \bar{U}^{-1}$, we can thus write:

$$\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)] = \frac{(\bar{U}^{-1})_k \text{var}(\bar{z}^*) (\bar{U}^{-1})_k'}{(\bar{U}^{-1})_k [n^{-1} S_z^*] (\bar{U}^{-1})_k'}, \quad (7)$$

where $(\bar{U}^{-1})_k$ is the k^{th} row of \bar{U}^{-1} . This simplifies in the case $q = 1$ to:

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = \text{var}(\bar{z}^*) / [n^{-1} S_z^*]. \quad (8)$$

We may explore more specific forms of these expressions under different models and assumptions about the weights and the sampling scheme. We focus here on the impact of clustering, assuming equal weights and no stratification. Consider the three-level model in (5) and, to simplify matters, suppose that $q = 1$ and $x_{ait} \equiv 1$ and β is the mean of y_{ait} . Then, straightforward algebra shows that the value of z_i^* for individual i within cluster a is $[1 + \rho^*(T-1)]^{-1} \sum_t (\eta_a + u_{ait} + v_{ait})$. Now suppose that two-stage sampling is employed with a common sample size m per cluster. Then, evaluating the variance $\text{var}(\bar{z}^*)$ and probability limit S_z^* in (8) with respect to the model in

(5), we find, in a similar manner to Skinner (1989a, page 38):

$$\text{meff}[\hat{\beta}, v_0(\hat{\beta})] = 1 + (m-1)\tau, \quad (9)$$

where $\tau = \sigma_{\eta_a}^2 / (\sigma_{\eta_a}^2 + \sigma_u^2 + \sigma_v^2 / T)$ is the intracluster correlation of z_i^* . We see that, under this model, the *meff* increases as T increases (provided $\sigma_v^2 > 0$) and thus the impact of clustering on variance estimation is greater in the longitudinal case than for the cross-sectional problem (where $T = 1$).

This finding depends on the rather strong assumption that the cluster effects η_a are constant over time. In fact, (9) still holds if we replace η_a by a time-varying effect η_{at} , provided we replace τ by $\tau = \text{var}(\bar{\eta}_a) / [\text{var}(\bar{\eta}_a) + \sigma_u^2 + \sigma_v^2 / T]$, where $\bar{\eta}_a = \sum_t \eta_{at} / T$. Now, the *meff* will increase as T increases if (and only if) $\sigma_u^2 + \sigma_v^2 / T$ decreases faster with T than $\text{var}(\bar{\eta}_a)$. Whether this is the case will depend on the particular application. However, we suggest that for many longitudinal surveys of individuals with area-based clusters (the kind of setting we have in mind), this condition is plausible. In such applications we may often expect σ_v^2 to be large relative to σ_u^2 (*i.e.*, for the cross-sectional intracluster correlation to be small) in particular as a result of wave-specific measurement error and thus for $\sigma_u^2 + \sigma_v^2 / T$ to decrease fairly rapidly as T increases. The socio-economic characteristics of areas may often be expected to be more stable and only in unusual situations might we expect measurement error to lead to much occasion-specific variance in η_{at} . Thus, we suggest that the ratio of $\text{var}(\bar{\eta}_a)$ for $T = 5$, say, compared to $T = 1$ may in such applications usually be expected to be greater than $(\sigma_u^2 + \sigma_v^2 / 5) / (\sigma_u^2 + \sigma_v^2)$ which will approach 1/5 as σ_u^2 / σ_v^2 approaches 0. We thus suggest that in many practical circumstances it will be more important to allow for clustering in longitudinal analyses than in corresponding cross-sectional analyses. An empirical illustration is provided in Section 4.

We now consider the properties of variance estimators based upon the three-level model in (5). We consider only the approach based upon the assumption of normally distributed homoscedastic random effects, ignoring survey weights, given the (virtual) equivalence of the 'robust' multilevel approach and linearization.

If model (5) is correct and we can indeed ignore survey weights then the model-based variance estimator will be consistent (Goldstein 1986). However, as discussed in Skinner (1989b, page 68) and supported by theory in Skinner (1986), the main feature of clustering likely to impact on the standard errors of estimated regression coefficients is the variation in regression coefficients between clusters. This is not allowed for in model (5).

To see how model (5) may fail to capture the effects of clustering adequately, consider the cross-sectional case ($T = 1$) where x is scalar. Then, if the three-level model (5) holds, an approximate expression for the meff of the variance estimator of $\hat{\beta}$ based upon the two-level model (2) is:

$$\text{meff} = 1 + (m - 1)\tau_1\tau_x, \quad (10)$$

where $\tau_1 = \sigma_\eta^2 / (\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$ and τ_x is the intraclass correlations for x (Scott and Holt 1982; Skinner 1989b, page 68). This result extends in the longitudinal case, to:

$$1 \leq \text{meff} \leq 1 + (m - 1)\bar{\tau}\tau_x, \quad (11)$$

where $\bar{\tau}$ is the long-run ($T = \infty$) version of τ (see Appendix) and τ_x is an intraclass correlation coefficient for $z_{ai} = \sum_t x_{ait} / T$. The proof of this result and the simplifying assumptions required are sketched in the Appendix. The main point is that both $\bar{\tau}$ and τ_x will often be small in which case $\bar{\tau}\tau_x$ will be very small and thus meff may be implausibly close to one with the model-based variance estimator being subject to downward bias. We explore this empirically in Section 4. Of course, random coefficients could be introduced into model (5) and we consider this also in Section 4. However, given the difficulty of specifying a correct random coefficient model, this approach does not seem likely to be very robust.

Our focus in this section has so far been on the potential bias (or inconsistency) of variance estimation methods. It is also desirable to consider their efficiency. In particular, the linearization method may be expected to be less efficient than model-based variance estimation if the model is correct. The relative importance of efficiency vs. bias may be expected to increase as the number of clusters decreases. Wolter (1985, Chapter 8) summarises a number of simulation studies investigating both the bias and variance of the linearization variance estimator and these studies suggest that the linearization method performs well even with few clusters. Possible degrees of freedom corrections to confidence intervals for regression coefficients based upon the linearization method with small numbers of clusters are discussed by Fuller (1984). A simulation study of estimators for multilevel models in Maas and Hox (2004) does not suggest that the linearization method performs noticeably worse than the model-based approach, in terms of the coverage of confidence intervals for coefficients in β , even with as few as 30 clusters.

4. Example: Regression analysis of BHPS data on attitudes to gender roles

We now present an application to BHPS data to illustrate some of the theoretical properties discussed in the previous section.

Recent decades have witnessed major changes in the roles of men and women in the family in many countries. Social scientists are interested in the relation between changing attitudes to gender roles and changes in behaviour, such as parenthood and labour force participation (e.g., Morgan and Waite 1987; Fan and Marini 2000). A variety of forms of statistical analysis are used to provide evidence about these relationships. Here, we consider estimating a linear model of form (1), with a measure of attitude to gender roles as the outcome variable, y , following an analysis of Berrington (2002).

The data come from waves 1, 3, 5, 7 and 9 (collected in 1991, 1993, 1995, 1997, and 1999 respectively) of the BHPS and these waves are coded $t = 1, \dots, T = 5$ respectively. Respondents were asked whether they 'strongly agreed', 'agreed', 'neither agreed nor disagreed', 'disagreed' or 'strongly disagreed' with a series of statements concerning the family, women's roles, and work out of the household. Responses were scored from 1 to 5. Factor analysis was used to assess which statements could be combined into a gender role attitude measure. The attitude score, y_{it} , considered here is the total score for six selected statements for woman i at wave t . Higher scores signify more egalitarian gender role attitudes. Berrington (2002) provides further discussion of this variable. A more sophisticated analysis might include a measurement error model for attitudes (e.g., Fan and Marini 2000), with each of the five-point responses to the six statements treated as ordinal variables. Here, we adopt a simpler approach, treating the aggregate score y_{it} and the associated coefficient vector β as scientifically interesting, with the measurement error included in the error term of the model.

Covariates for the regression analysis were selected on the basis of discussion in Berrington (2002) but reduced in number to facilitate a focus on the methodological issues of interest. The covariate of primary scientific interest is economic activity, which distinguishes in particular between women who are at home looking after children (denoted 'family care') and women following other forms of activity in relation to the labour market. Variables reflecting age and education are also included since these have often been found to be strongly related to gender role attitudes (e.g., Fan and Marini 2000). All these covariates may change values between waves. A year variable (scored 1, 3, ..., 9) is also included. This may reflect both historical change and the general ageing of the women in the sample.

The BHPS is a household panel survey of individuals in private domiciles in Great Britain (Taylor, Brice, Buck and Prentice-Lane 2001). The initial (wave one) sample in 1991 was selected by a stratified multistage design in which households had approximately equal probabilities of inclusion. The households were clustered into 250 primary

sampling units (PSUs), consisting of postcode sectors. All resident members aged 16 or over were selected in sample households. All adults selected at wave one were followed from wave two onwards and represent the longitudinal sample. The survey is subject to attrition and other forms of wave non-response. To handle this non-response, we have simply replaced s in (3) by the ‘longitudinal sample’ of individuals for which observations are available for each of $t = 1, \dots, T$ and have chosen not to apply any survey weighting since our aim is to study potential misspecification effects associated with clustering and we wish to avoid confounding these with weighting effects. We also ignore the impact of stratification in the numerical work in this section (but see Section 5 for some comments on the effect of weights and stratification).

Given the analytic interest in whether women’s primary labour market activity is ‘caring for a family’, we define our study population as women aged 16-39 in 1991. Thus our data consist of the longitudinal sample of women in the eligible age range for whom full interview outcomes (complete records) were obtained in all five waves, a sample of $n = 1,340$ women. These women are spread fairly evenly across 248 postcode sectors. The small average sample size of around five per postcode sector combined with the relatively low intra-postcode sector correlation for the attitude variable of interest leads to relatively small impacts of the design, as measured by meffs. Since our aims are methodological ones, we have chosen to group the postcode sectors into 47 geographically contiguous clusters, to create sharper comparisons, less blurred by sampling errors which can be appreciable in variance estimation. The meffs in the tables we present therefore tend to be greater than they are for the actual design. The latter results tend to follow similar patterns, although the patterns are less clear-cut as a result of sampling error.

We first estimate meffs for the linearization estimator, as discussed at the beginning of Section 3. Using data from just the first wave and setting $x_{ait} \equiv 1$, the estimated meff for this cross-sectional mean is given in Table 1 as about 1.5. This value is plausible since, if we make the usual approximation of (9) for unequal sample cluster sizes by replacing m by \bar{m} , the average sample size per cluster, we find that $1 + (\bar{m} - 1)\tau = 1.5$ and $\bar{m} = 1,340/47 \approx 29$ imply a value of τ of about 0.02 and such a small value is in line with other estimated values of τ found for attitudinal variables in British surveys (Lynn and Lievesley 1991, Appendix D).

To assess the impact of the longitudinal aspect of the data, we estimated a series of meffs using data for waves $1, \dots, t$ for $t = 2, 3, \dots, 5$. Although these estimated meffs are subject to sampling error, there seems clear evidence in Table 1 of a tendency for the meff to increase with the number of waves. This trend might be anticipated from the theoretical discussion in Section 3 if the average level of egalitarian attitudes in an area varies less from year to year than the attitude scores of individual women. This seems plausible since the latter will be affected both by measurement error and genuine changes in attitudes, so that $\text{var}(\bar{\eta}_a)$ may be expected to decline more slowly with T than $\text{var}(\bar{u}_a + \bar{v}_a)$. We may therefore expect τ , and consequently the meff, to increase as T increases, as we observe in Table 1.

We next elaborate the analysis by including indicator variables for economic activity as covariates. The resulting regression model has an intercept term and four covariates representing contrasts between women who are employed full-time and women in other categories of economic activity. The estimated meffs are presented in Table 2. The intercept term is a domain mean and standard theory for a meff of a mean in a domain cutting across clusters (Skinner 1989b, page 60) suggests that it will be somewhat less than the meff for the mean in the whole sample, as indeed is observed with the meff for the cross-section domain mean of 1.13 in Table 2 being less than the value 1.51 in Table 1. As before, there is some evidence in Table 2 of tendency for the meff to increase, from 1.13 with one wave to 1.50 with five waves, albeit with lower values of the meffs than in Table 1. The meffs for the contrasts in Table 2 vary in size, some greater than and some less than one. These meffs may be viewed as a combination of the traditional variance inflating effect of clustering in surveys together with the variance reducing effect of blocking in an experiment. Such variance reduction arises if the domains being contrasted share a common cluster effect (of the form η_a in model (5)) which tends to cancel out in the contrasts, implying that the actual variance of the contrast is lower than the expectation of the variance estimator which assumes independence between domains. The latter expectation will be inflated by common cluster effects. The main feature of these results of interest here is that there is again no tendency for the meffs to converge to one as the number of waves increases. If there is a trend, it is in the opposite direction. For the contrast of particular scientific interest, that between women who are full-time employed and those who are ‘at home caring for a family’, the meff is consistently well below one.

We next refine the model further by including, as additional covariates, age group, year and qualifications. The estimated meffs are given in Table 3. The meffs for the regression coefficients corresponding to categories of

Table 1 Estimates for longitudinal means

	$\hat{\beta}$	s.e.	meffs				
Waves	1-9	1-9	1	1,3	1,3,5	1-7	1-9
	19.83	0.12	1.51	1.50	1.68	1.81	1.84

economic activity again vary, some being above one and some below one, for the same reasons as for the contrasts (which may also be interpreted as regression coefficients) in Table 2. There is again some evidence of a tendency for these meffs to diverge away from one as the number of waves increases. A comparison of Tables 1 and 3 confirms the observation of Kish and Frankel (1974) that meffs for regression coefficients tend not to be greater than meffs for the means of the dependent variable.

Table 2 Estimates for regression with covariates defined by economic activity

	$\hat{\beta}$ s.e.		meffs					
Waves	1-9	1-9	1	1,3	1,3,5	1-7	1-9	
Intercept	20.58	0.11	1.13	1.01	1.09	1.38	1.50	
Contrasts for								
PT employed	-1.03	0.10	0.93	0.91	0.93	1.00	0.89	
Other inactive	-0.80	0.15	0.60	0.96	0.68	0.76	0.81	
FT student	0.41	0.24	1.10	1.32	1.14	1.48	1.44	
Family care	-2.18	0.10	0.72	0.49	0.58	0.66	0.60	

Note: a) intercept is mean for women full-time employed
b) contrasts are for other categories of economic activity relative to full-time employed

Table 3 Estimates for regression coefficients with additional covariates in model

	$\hat{\beta}$ s.e.		meffs					
Waves	1-9	1-9	1	1,3	1,3,5	1-7	1-9	
Intercept	20.20	0.30	0.95	0.87	0.87	1.04	1.07	
Year, <i>t</i>	-0.04	0.01	-	0.86	0.69	0.59	0.96	
Age Group								
16-21	0.00	-						
22-27	-0.71	0.25	1.22	1.37	1.44	1.73	1.64	
28-33	-0.89	0.27	1.38	1.40	1.46	1.68	1.59	
34+	-1.03	0.27	0.94	1.10	1.13	1.26	1.34	
Economic Activity								
FT employed	0.00	-						
PT employed	-0.93	0.10	0.97	0.95	0.96	1.06	0.91	
Other inactive	-0.75	0.15	0.60	0.96	0.68	0.77	0.81	
FT student	0.17	0.24	0.93	1.32	1.23	1.39	1.32	
Family care	-2.09	0.10	0.77	0.59	0.70	0.78	0.67	
Qualification								
Degree	0.00	-						
QF	-0.52	0.21	0.77	0.64	0.75	0.87	0.85	
A-level	-0.61	0.24	0.98	0.87	0.94	0.94	1.01	
O-level	-0.44	0.20	0.62	0.62	0.59	0.69	0.73	
Other	-1.16	0.22	0.83	0.83	0.78	0.80	0.82	

We next consider model-based standard errors obtained from the three level model in (5), as discussed in section 2. The results are given in Table 4 in the column headed ‘3

level model-based’. For comparison, we also estimate the standard errors under the two level model in (2) - the results are in the column headed ‘2 level model-based’. The estimates in the two columns are virtually identical. There is a single digit difference in the third decimal place for some coefficients and slightly greater difference for the intercept term. We suggest that this is evidence that simply adding in a random area effect term can seriously understate the impact of clustering on the standard errors of the estimated regression coefficients. This evidence is in line with the theoretical upper bound for the meff in (11). The estimated value of $\bar{\tau}$ in (11) is 0.019 and none of the covariates may be expected to display important intra-area correlation so the expected values of the variance estimators for the two-level and three-level models would be expected to be very close.

We suggested in Section 3 that the main feature of clustering likely to impact on the covariance matrix of $\hat{\beta}$ is the variation in regression coefficients between clusters. We have explored this idea by introducing random coefficients in the model. Treating the elements of β now as the expected values of the random coefficients, we found that the estimates of β were hardly changed. We found that the estimated standard errors of these estimates were indeed inflated, much more so than from the introduction of the extra cluster random effect in model (5), and that the inflation was of an order similar to those of the meffs in Tables 2 and 3. Nevertheless, the IGLS method did lead to several negative estimates of the variances of the random coefficients, raising issues of which coefficients to allow to vary or more generally the issue of model specification. This problem is accentuated with increasing numbers of covariates, as the number of parameters in the covariance matrix of the coefficient vector increases with the square of the number of covariates. Overall, the inclusion of random coefficients seems to raise at least as many problems as it solves, if the clustering is not of intrinsic scientific interest, and thus does not seem a very satisfactory way to allow for clustering in variance estimation. It is simpler to change the method of variance estimation.

As mentioned at the end of Section 2, one alternative is a ‘robust’ variance estimation method based on the model in (5) (Goldstein 2003, page 80). Values of such robust standard error estimates are also included in Table 4. As anticipated in Section 2, the robust standard error estimator for the two level model performs very similarly to the linearization estimator which ignores clustering. The robust standard error estimator for the three level model performs very similarly to the linearization estimator which allows for two stage sampling. The slight differences reflect the differences between the methods of estimating V .

Table 4 Estimated standard errors of regression coefficients

	Linearization		Multilevel modelling			
	SRS	complex	2 level model-based	2 level robust	3 level model-based	3 level robust
Intercept	0.287	0.296	0.253	0.288	0.259	0.293
Year, t	0.014	0.014	0.013	0.014	0.013	0.014
Age Group						
16-21						
22-27	0.191	0.245	0.155	0.192	0.155	0.243
28-33	0.214	0.270	0.187	0.215	0.187	0.266
34+	0.237	0.275	0.218	0.238	0.218	0.271
Economic Activity						
FT employed						
PT employed	0.103	0.098	0.098	0.103	0.098	0.096
Other inactive	0.166	0.150	0.146	0.166	0.146	0.148
FT student	0.207	0.238	0.199	0.207	0.199	0.236
Family care	0.125	0.102	0.112	0.125	0.112	0.101
Qualification						
Degree						
QF	0.228	0.210	0.207	0.228	0.208	0.211
A-level	0.238	0.239	0.209	0.240	0.210	0.237
O-level	0.234	0.199	0.217	0.235	0.218	0.199
Other	0.247	0.224	0.229	0.249	0.230	0.223

The linearization method in the presence of two-stage sampling is thus very close to robust variance estimation methods used in the literature on multilevel modeling. The distinction between the methods becomes stronger if we allow also for stratification and weighting. Another distinction is that in the multilevel modeling approach, differences between model-based and the robust standard errors might be used as a diagnostic tool to detect departures from the model (Maas and Hox 2004). For example, the large differences in the three-level standard errors for the coefficients of age group in Table 4 might lead to consideration of the inclusion of random coefficients for age group. This contrasts with the survey sampling approach where the error structure in model (5) is only treated as a working model and it is not necessarily expected that standard errors based upon this model will be approximately valid.

5. Discussion

We have presented some theoretical arguments and empirical evidence that the impact of ignoring clustering in standard error estimation for certain longitudinal analyses can tend to be larger than for corresponding cross-sectional analyses. The implication is that it is, in general, at least as important to allow for clustering in standard errors for longitudinal analyses as for cross-sectional analyses and that the findings of, for example, Kish and Frankel (1974),

should not be used as grounds to ignore complex sampling in the former case.

The longitudinal analyses considered in this paper are of a certain kind and we should emphasise that the patterns observed for meffs in these kinds of analyses may well not extend to all kinds of longitudinal analyses. To speculate about the class of models and estimators for which the patterns observed in this paper might apply, we conjecture that increased meffs for longitudinal analyses will arise when the longitudinal design enables temporal 'random' variation in individual responses to be extracted from between-person differences and hence to reduce the component of standard errors due to these differences, but provides less 'explanation' of between cluster differences, so that the relative importance of this component of standard errors becomes greater.

The empirical work presented in this paper has also been restricted to the impact of clustering. We have undertaken corresponding work allowing for weighting and stratification and found broadly similar findings. Stratification tends to have a smaller effect than clustering. The sample selection probabilities in the BHPS do not vary greatly and the impact of weighting by the reciprocals of these probabilities on both point and variance estimates tends not to be large. There is rather greater variation among the longitudinal weights which are provided with BHPS data for analyses of sets of individuals who have responded at each wave up to and including a given year T . The impact

of these weights on point and variance estimates is somewhat greater. As T increases and further attrition occurs, the longitudinal weights tend to become more variable and lead to greater inflation of variances. This tends to compound the effect we have described of meffs increasing with T .

Leaving aside consideration of stratification and weighting, we have compared two approaches to allowing for cluster sampling. We have treated the survey sampling approach as a benchmark. We have also considered a multilevel modelling approach to allow for clustering. We have suggested that the use of a simple additive random effect to represent clustering can seriously understate the impact of clustering and may lead to underestimation of standard errors. If the clustering is of scientific interest, one solution would be to consider including random coefficients. Another would be to use the 'GEE2' approach (Liang, Zeger and Qaqish 1992) and specify an additional parametric model for $E(y_i | y'_i)$. If the clustering is treated as a nuisance, simply reflecting administrative convenience in data collection, we suggest the survey sampling approach has a number of practical advantages. This is discussed further by Lavange *et al.* (1996, 2001) in relation to other applications to repeated measures data.

Appendix

Justification for (11)

For simplicity, x and β are taken to be scalar, $\hat{\beta}$ is taken to be the ordinary least squares estimator and it is assumed that the sample sizes within clusters are all equal to m . The meff in (11) is defined as $\text{var}_3(\hat{\beta})/E_3[v_2(\hat{\beta})]$, where E_3 and var_3 are moments with respect to the three-level model in (5) and $v_2(\hat{\beta})$ is a variance estimator based upon the two-level model in (2). Under (5) we obtain

$$\text{var}_3(\hat{\beta}) = \left(\sum_{cit} x_{cit}^2 \right)^{-2} \left(\sigma_\eta^2 \sum_c x_{c++}^2 + \sigma_u^2 \sum_{ci} x_{ci+}^2 + \sigma_v^2 \sum_{cit} x_{cit}^2 \right),$$

where $+$ denotes summation across a suffix, σ_η^2 , σ_u^2 and σ_v^2 are the respective variances of η_a , u_{ai} and v_{ait} and x_{cit} is centred at 0. We further suppose that $v_2(\hat{\beta})$ is defined so that $E[v_2(\hat{\beta})] \approx (\sum_{cit} x_{cit}^2)^{-2} [(\sigma_\eta^2 + \sigma_u^2) \sum_{ci} x_{ci+}^2 + \sigma_v^2 \sum_{cit} x_{cit}^2]$. After some algebra we may show that

$$\text{meff} = 1 + (m-1) \tilde{\tau} \tau_x \rho [1 + (T-1) \tau_x] / [1 + (T-1) \rho \tau_x], \quad (12)$$

where $\tilde{\tau} = \sigma_\eta^2 / (\sigma_\eta^2 + \sigma_u^2)$, $\rho = (\sigma_\eta^2 + \sigma_u^2) / (\sigma_\eta^2 + \sigma_u^2 + \sigma_v^2)$, $\tau_x = \sigma_{xB}^2 / \sigma_x^2$, $\sigma_x^2 = \sum_{cit} x_{cit}^2 / (nT)$, $\sigma_{xB}^2 = [\sum_{ci} (x_{ci+} / T)^2 / n - \sigma_x^2 / T] / [1 - 1/T]$, $\tau_z = \sigma_{zB}^2 / \sigma_z^2$, $\sigma_z^2 = \sum_{ci} z_{ci}^2 / n$, $\sigma_{zB}^2 = [\sum_{ci} (z_{ci} / m)^2 / C - \sigma_z^2 / m] / [1 - 1/m]$ and $n = Cm$ is the sample size. Note that $\tilde{\tau} \rho = \tau_1$ and, when

$T = 1$, $\tau_z = \tau_x$ so that (12) reduces to (10). In general $\rho \leq 1$ and (11) follows from (12). In fact, we estimate ρ as 0.59 in our application so the bound in (11) is not expected to be very tight.

Acknowledgements

The research of the second author was supported by grant 20.0286/01.3 from the Brazilian National Council for Scientific and Technological Development (CNPq).

References

- Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*. 2nd Ed. Chichester: John Wiley & Sons, Inc.
- Berrington, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study. In *Meaning and Choice: Value Orientations and Life Course Decisions*, (Ed., R. Lesthaeghe) Brussels: NIDI.
- Chambers, R.L., and Skinner, C.J. Eds. (2003). *Analysis of Survey Data*. Chichester: John Wiley & Sons, Inc.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-92.
- Diggle, P.J., Heagerty, P., Liang, K. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2nd Ed. Oxford: Oxford University Press.
- Fan, P.-L., and Marini, M.M. (2000). Influences on gender-role attitudes during the transition to adulthood. *Social Science Research*, 29, 258-283.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*. Vol. 37, Series C, 117-132.
- Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.
- Fuller, W.A., and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 74, 430-431.
- Goldstein, H. (2003). *Multilevel Statistical Models*, 3rd Ed. London: Arnold.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.
- Lavange, L.M., Koch, G.G. and Schwartz, T.A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, 20, 2609-23.

- Lavage, L.M., Stearns, S.C., Lafata, J.E., Koch, G.G. and Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, 5, 311-329.
- Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K.Y., Zeger, S.L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, 54, 3-40.
- Lipsitz, S.R., Fitzmaurice, G.M., Orav, E.J. and Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270-278.
- Lynn, P., and Lievesley, D. (1991). *Drawing General Population Samples in Great Britain*. London: Social and Community Planning Research.
- Maas, C.J.M., and Hox, J.J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427-440.
- Morgan, S.P., and Waite, L.J. (1987). Parenthood and the attitudes of young adults. *Am. Sociological Review*, 52, 541-547.
- Pfiffermann, D., Skinner, C., Holmes, D., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-56.
- Renard, D., and Molenberghs, G. (2002). Multilevel modelling of complex survey data. In *Topics in Modelling Clustered Data* (Eds., M. Aerts, H. Geys, G. Molenberghs and L.M. Ryan). Boca Raton: Chapman and Hall/CRC. 263-272.
- Scott, A.J., and Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.
- Shah, B.V., Barnwell, B.G. and Bieler, G.S. (1997). SUDAAN User's manual, release 7.5. Research triangle park, NC: Research Triangle Institute.
- Skinner, C.J. (1986). Design effects of two stage sampling. *Journal of the Royal Statistical Society, Series B*, 48, 89-99.
- Skinner, C.J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt and T.M.F. Smith) Chichester: John Wiley & Sons, Inc. 23-58.
- Skinner, C.J. (1989b). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt and T.M.F. Smith) Chichester: John Wiley & Sons, Inc. 59-87.
- Skinner, C.J., Holt, D. and Smith, T.M.F. Eds. (1989). *Analysis of Complex Surveys*. Chichester: John Wiley & Sons, Inc.
- Taylor, M.F. ed, Brice, J., Buck, N. and Prentice-Lane, E. (2001). *British Household Panel Survey - User Manual - Volume A: Introduction, Technical Report and Appendices*. Colchester, University of Essex.
- Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer.

Modelling durations of multiple spells from longitudinal survey data

Milorad S. Kovačević and Georgia Roberts¹

Abstract

We investigate some modifications of the classical single-spell Cox model in order to handle multiple spells from the same individual when the data are collected in a longitudinal survey based on a complex sample design. One modification is the use of a design-based approach for the estimation of the model coefficients and their variances; in the variance estimation each individual is treated as a cluster of spells, bringing an extra stage of clustering into the survey design. Other modifications to the model allow a flexible specification of the baseline hazard to account for possibly differential dependence of hazard on the order and duration of successive spells, and also allow for differential effects of the covariates on the spells of different orders. These approaches are illustrated using data from the Canadian Survey of Labour and Income Dynamics (SLID).

Key Words: Cox regression; Design-based inference; Model-based inference; Spell order; SLID.

1. Introduction

The modelling problem addressed in this paper is known under different names such as correlated failure-time modelling, multivariate survival modelling, multiple spells modelling, or a recurrent events problem. It has been studied in the biomedical (*e.g.*, Lin 1994, Hougaard 1999), social (Blossfeld and Hamerle 1989, Hamerle 1989) and economic literature (Lancaster 1979, Heckman and Singer 1982). Generally this type of modelling is required to address issues that arise in time-to-event studies when two or more events occur to the same subject and where the research interest is to assess the effect of various covariates on the length of a spell. Failure times are correlated within a subject, and thus the assumption of independence of failure times conditional on given measured covariates, required by standard survival models, is likely to be violated. In studies of duration of spells (poverty, unemployment, *etc.*), the “failure” is equivalent to exiting from the state of interest. An additional property of many multiple spells, often ignored, is that the spells are ordered “events”; that is, the second spell cannot occur before the first, *etc.* This paper was motivated by a study of unemployment spells, discussed further in Section 5.

The dependence among the spells from the same individual arises from the fact that these spells share certain unobserved characteristics of the individual. The effect of these unobserved characteristics can be explicitly modelled as a random effect (*e.g.*, Clayton and Cuzick 1985). When this is done, it is assumed that the random effect follows a known statistical distribution. The gamma distribution with mean 1 and unknown variance is the distribution of choice in many applications. Then, estimates of random and fixed

effects can be obtained by some suitable method (*e.g.*, two-stage likelihood (Lancaster 1979), using an EM algorithm (Klein 1992), *etc.*). This paper does not explore this approach.

Another approach that has been taken - and is the one that we will be using - is to take a semi-parametric approach where we do not explicitly model the dependence among multiple spells. We model the marginal distributions of the individual spells, with a possible utilization of the order of the spells in the model specification. In the non-survey context, Lin (1994) describes how it is sufficient just to modify the “naïve” covariance matrix of the estimated model coefficients obtained under the assumption of independence since the correlated durations need to be accounted for in the variance estimates but not in the estimates of coefficients per se.

In socio-economic studies of spell durations the data sources are frequently longitudinal surveys with complex sample designs that involve stratification, sampling in several stages, selection with unequal probabilities, stochastic adjustments for attrition and non-response, calibration to known parameters, *etc.* Consequently, it is necessary to account for the impact of the sample design on the distribution of the sample data when estimating model parameters and the variances of these estimates. Our approach when analyzing complex survey data is to model the marginal distributions of the multiple spells using single-spell methods, treating the dependence among the spells as a nuisance - both the dependence due to the correlation of spells from the same person and dependence among individuals due to the survey design - but taking account of the unequal selection probabilities through the survey weights. Based on the model chosen, finite population

1. Milorad S. Kovačević, Methodology Research Advisor, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: kovamil@statcan.ca; Georgia Roberts, Chief of the Data Analysis Resource Center at the Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6. E-mail: robertg@statcan.ca.

parameters are defined and estimated as in Binder (1992). Standard errors are estimated using an appropriate design-consistent linearization method under the assumption that the primary sampling units are sampled with replacement within strata. This assumption is viable when the sampling rates at the first stage are small, as is generally the case in socio-economic surveys. Also, for such samples, the difference between finite population and superpopulation inference (*i.e.*, the standard errors and the test statistics) has been found to be rather negligible (Lin 2000). Therefore, the results from inference based on our approach extend beyond the finite population under study.

In the next section we review single-spell modelling and some methods for robust estimation of variances when the model is misspecified - first under a model-based framework and then under a design-based one. Section 3 contains further discussion of robust variance estimation for multiple spells. In Section 4, we introduce three models for multiple spells and describe how to fit these models using design-based robust estimation methods. In Section 5, we fit these models to data from the Canadian Survey of Labour and Income Dynamics (SLID) and discuss the results. Finally, Section 6 contains some overall remarks.

2. Inference for the single-spell hazard rate model

The duration of a spell (or simply, a spell) experienced by an individual is a random variable denoted by T . We are particularly interested in the hazard function $h(t)$ of T at time t , defined as the instantaneous rate of spell completion at time t given that it has not been completed prior to time t , formally

$$h(t) = \lim_{dt \rightarrow 0} \frac{\text{Prob}\{t \leq T < t + dt \mid T \geq t\}}{dt}.$$

The value of the hazard function at t is called the exit rate to emphasize that the completion of the spell is equivalent to exiting the state of interest. Duration models and analysis of duration in general are formulated and discussed in terms of the hazard function and its properties.

From a subject matter perspective, frequently the main interest is to study the impact of some key covariates on the distribution of T . A proportional hazards model is often chosen for such a study. Under the proportional hazards model, the hazard function of the spell T given a vector of possibly time-varying covariates $\mathbf{x}(t) = (x_1(t), \dots, x_p(t))'$ is

$$h(t \mid \mathbf{x}(t)) = \lambda_0(t) e^{\mathbf{x}'(t)\boldsymbol{\beta}}, \quad (1)$$

The function $\lambda_0(t)$ is an unspecified baseline hazard function and gives the shape of $h(t \mid \mathbf{x}(t))$. The baseline hazard describes the duration dependence, such as whether

the hazard rate depends on time already spent in the spell. For example, negative dependence describes the situation where the longer the spell the smaller the probability of exit. If an individual has all $\mathbf{x}(t)$ variables set at 0, the value (level) of the hazard function is equal to the baseline hazard.

2.1 Model-based inference

The vector $\boldsymbol{\beta}$ contains the unknown regression parameters showing the dependence of the hazard on the $\mathbf{x}(t)$ vector, and may be estimated by maximizing the partial likelihood function (Cox 1975):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \left[\frac{e^{\mathbf{x}'_i(T_i)\boldsymbol{\beta}}}{\sum_{j=1}^n Y_j(T_i) e^{\mathbf{x}'_j(T_i)\boldsymbol{\beta}}} \right]^{\delta_i}. \quad (2)$$

Here T_1, \dots, T_n are n possibly right-censored durations; $\delta_i = 1$ if T_i is an observed duration and $\delta_i = 0$ otherwise; and $\mathbf{x}_i(t)$ is the corresponding covariate vector observed on $[0, T_i]$. The denominator sum is taken over the spells that are at risk of being completed at time T_i , *i.e.*, $Y_j = 1$ if $t \leq T_j$, and is equal to 0 otherwise. The estimate $\hat{\boldsymbol{\beta}}$ of the model parameter $\boldsymbol{\beta}$ is obtained by solving the partial likelihood score equation

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^n u_{i0}(T_i, \boldsymbol{\beta}) = 0, \quad (3)$$

where

$$u_{i0}(T_i, \boldsymbol{\beta}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{S^{(1)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} \right\}, \quad (4)$$

$$S^{(0)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) e^{\mathbf{x}'_i(t)\boldsymbol{\beta}}, \quad (5)$$

and

$$S^{(1)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{x}_i(t) e^{\mathbf{x}'_i(t)\boldsymbol{\beta}}. \quad (6)$$

If model (1) is true and the durations are independent, the model-based variance matrix of the score function $U_0(\boldsymbol{\beta})$ is

$$J(\boldsymbol{\beta}) = -\partial U_0(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \sum_{i=1}^n \delta_i \left\{ \frac{S^{(2)}(T_i, \boldsymbol{\beta})}{S^{(0)}(T_i, \boldsymbol{\beta})} - \frac{S^{(1)}(T_i, \boldsymbol{\beta})[S^{(1)}(T_i, \boldsymbol{\beta})]'}{[S^{(0)}(T_i, \boldsymbol{\beta})]^2} \right\},$$

where

$$S^{(2)}(t, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n Y_i(t) \mathbf{x}_i(t) \mathbf{x}'_i(t) e^{\mathbf{x}'_i(t)\boldsymbol{\beta}}.$$

The approximate variance of $\hat{\beta}$, obtained by linearization, is $J^{-1}(\hat{\beta})$.

If the form of (1) is incorrect but observations are independent, Lin and Wei (1989) provide the robust variance estimator for $\hat{\beta}$ as

$$J^{-1}(\hat{\beta}) G(\hat{\beta}) J^{-1}(\hat{\beta}), \quad (7)$$

where

$$G(\beta) = \sum_{i=1}^n g_i(\beta) g_i'(\beta)$$

and

$$g_i(\beta) = u_{i0}(T_i, \beta)$$

$$-\sum_{j=1}^n \delta_j \frac{Y_j(T_j) e^{x_j'(T_j)\beta}}{n S^{(0)}(T_j, \beta)} \left\{ \mathbf{x}_j(T_j) - \frac{S^{(1)}(T_j, \beta)}{S^{(0)}(T_j, \beta)} \right\}. \quad (8)$$

2.2 Design-based inference

For observations from a survey with a complex sample design, Binder (1992) used a pseudo-likelihood method to estimate the parameters and their variances for a proportional hazards model in the case of a single spell per individual. In particular, he first defined the finite population parameter of interest as a solution of the partial likelihood score equation (3) calculated from the spells of the finite population targeted by the survey:

$$U_0(\mathbf{B}) = \sum_{i=1}^N u_{i0}(T_i, \mathbf{B}) = 0,$$

where $u_{i0}(T_i, \mathbf{B})$ is the score residual defined in the same way as $u_{i0}(T_i, \hat{\mathbf{B}})$, except that the averages in the definitions of $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ extend over N observations rather than n . Note that if all members of the finite population targeted by the survey do not experience spells, N represents the size of the subpopulation that experiences spells, and the summation is over these N individuals.

An estimate $\hat{\mathbf{B}}$ of the parameter \mathbf{B} is obtained as a solution to the partial pseudo-score estimating equation

$$\hat{U}_0(\hat{\mathbf{B}}) = \sum_{i=1}^N w_i(s) \hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

where $w_i(s) = w_i$, the survey weight, if $i \in s$, and 0 otherwise. Function $\hat{u}_{i0}(T_i, \hat{\mathbf{B}})$ takes the form

$$\hat{u}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i \left\{ \mathbf{x}_i(T_i) - \frac{\hat{S}^{(1)}(T_i, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

where

$$\hat{S}^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^N w_i(s) Y_i(t) e^{x_i'(t)\hat{\mathbf{B}}},$$

and

$$\hat{S}^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^N w_i(s) Y_i(t) \mathbf{x}_i(t) e^{x_i'(t)\hat{\mathbf{B}}}.$$

Generally, the design-based variance of an estimate $\hat{\theta}$ that satisfies an estimating equation of the form $\hat{U}(\hat{\theta}) = \sum w_i u_i(\hat{\theta}) = 0$ can be estimated, using linearization, as

$$\hat{J}^{-1} \hat{V}(\hat{U}(\hat{\theta})) \hat{J}^{-1}, \quad (9)$$

where $\hat{J} = \partial \hat{U}(\hat{\theta}) / \partial \theta$ is evaluated at $\theta = \hat{\theta}$, and $\hat{V}(\hat{U}(\hat{\theta}))$ is the estimated variance of the estimated total $\hat{U}(\hat{\theta})$ obtained by some standard design-based variance estimation method (see for example Cochran (1977)) and evaluated at $\theta = \hat{\theta}$. Binder (1983) states that in order to use this approach to derive a consistent estimate of the variance, $\hat{U}(\hat{\theta})$ must be expressed as a sum of independent random vectors. In the case of the proportional hazards model above, $\hat{U}_0(\hat{\mathbf{B}})$ does not satisfy this condition since each \hat{u}_{i0} is a function of $\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})$, both of which include many individuals besides the i^{th} one. Thus, Binder (1992) found an alternative expression for $\hat{U}_0(\hat{\mathbf{B}})$ which conforms to these conditions, making it possible to obtain a design consistent estimate $\hat{V}(\hat{U}_0(\hat{\mathbf{B}}))$ by application of a design-based variance estimation method to the alternate expression and then evaluating this variance estimate at $\mathbf{B} = \hat{\mathbf{B}}$. If the design-based variance estimation method chosen is the linearization method, then the first step consists of calculating the following residual for each of the sampled individuals:

$$\hat{u}_i(T_i, \hat{\mathbf{B}}) = \hat{u}_{i0}(T_i, \hat{\mathbf{B}})$$

$$-\sum_{j=1}^N w_j(s) \delta_j \frac{Y_j(T_j) e^{x_j'(T_j)\hat{\mathbf{B}}}}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \left\{ \mathbf{x}_j(T_j) - \frac{\hat{S}^{(1)}(T_j, \hat{\mathbf{B}})}{\hat{S}^{(0)}(T_j, \hat{\mathbf{B}})} \right\}. \quad (10)$$

Each individual in the sample belongs to a particular PSU within a particular stratum. Thus, instead of identifying an individual by a single subscript i we will use a triple subscript hci where $h = 1, 2, \dots, H$ identifies the stratum, $c = 1, 2, \dots, c_h$ identifies the PSU within the stratum and $i = 1, 2, \dots, n_{hc}$ identifies the individual within the PSU. Then

$$\hat{V}(\hat{U}_0(\hat{\mathbf{B}})) = \sum_{h=1}^H \frac{1}{c_h(c_h - 1)} \sum_{c=1}^{c_h} (t_{hc} - \bar{t}_h) (t_{hc} - \bar{t}_h)',$$

where

$$t_{hc} = c_h \sum_{i=1}^{n_{hc}} w_{hci} \hat{u}_{hci} \text{ and } \bar{t}_h = \frac{c_h}{c_h} \sum_{c=1}^{c_h} t_{hc} / c_h.$$

3. Inference for multiple-spell hazard rate models

3.1 Model-based inference

If more than one spell is observed for an individual, it is realistic to assume that these spells are not independent. Thus, the partial likelihood function (2) is misspecified for multiple spells since it does not account for intra-individual correlation of the spells observed on the same individual. Following Lin and Wei (1989), it is sufficient to modify only the covariance matrix of the estimated model parameters since the correlated durations affect the variance while the model parameters can be estimated consistently without accounting for this correlation. Lin (1994) demonstrates how the covariance matrix of the estimated model parameters might be estimated when there is intra-individual correlation of spells, provided that spells from different individuals are independent.

3.2 Design-based inference

In a longitudinal survey with a multi-stage design, the multiple events can be correlated at different levels: the spells are clustered within an individual, and individuals are clustered within high-stage units. The positive intraclass correlation at any level adds extra variation to estimates calculated from such data, beyond what is expected under independence. The assumption of independence of observations when they are cluster-correlated leads to underestimating the true standard errors, which inflates the values of test statistics, and ultimately results in too-frequent rejection of null hypotheses. Thus, for multiple spells for individuals, where the data are from a longitudinal survey, accounting just for correlation within individuals is insufficient.

Design-based variance estimation for nested cluster-correlated data can be greatly simplified when it is reasonable to assume that individuals from different primary sampling units (PSU's) are uncorrelated. This is equivalent to assuming that the PSU's are sampled with replacement. This assumption also holds approximately when the first stage units are obtained by sampling without replacement, provided that the sampling rate at the first stage is very small. In such a case, an estimate of the between-PSU variability captures the variability among units in all subsequent stages, regardless of the dependence structure among observations within each PSU. For a recent summary of robust variance estimation for cluster-correlated data see Williams (2000). This implies that Binder's (1992) approach for robust variance estimation of the single-spell

model in the case of a survey design having with-replacement sampling at the first stage can be directly applied to the multiple spell situation since it accounts for the impact of cluster-correlation at all levels within each PSU.

4. Three models for multiple spells

In order to allow the covariates to have different effects for spells of different orders, as well as to allow different time dependencies (baseline hazards), we are exploring three models for multiple spells. The models differ according to the definition of the risk set and the assumptions about the baseline hazard. Two of these models account for the order of the spells.

It should be noted, however, that in our work, spell order refers only to spells occurring in the observation period from which the data are collected and not to the entire history of an individual (unless these two time periods coincide). For example, by the first spell we mean a first spell in the observation period although it may be a spell of some higher absolute order over the person's lifetime. This limitation implies a careful interpretation of any impact that spell order may have on covariate effects or on time dependency.

Model 1: In the first model, the risk set is carefully defined to take spell order into account in the sense that an individual cannot be at risk of completing the second spell before he completes the first, *etc.* This model, known as the conditional risk set model, was proposed by Prentice, Williams and Peterson (1981) and was reviewed by Lin (1994). It was also discussed by Hamerle (1989) and Blossfeld and Hamerle (1989) in the context of modelling multi-episode processes. Generally, the conditional risk set at time t for the completion of a spell of order j consists of all individuals that are in their j^{th} spells. This model allows spell order to influence both the effect of covariates and the shape of the baseline hazard function.

The hazard function for the i^{th} individual for the spell of j^{th} order is

$$h_j(t | \mathbf{x}_{ij}(t)) = \lambda_{0j}(t) e^{\mathbf{x}_{ij}^{(j)}(t) \boldsymbol{\beta}_j},$$

where, for each spell order, a different baseline hazard function and a different coefficient vector are allowed. For this model and for other models that will be considered in this Section, time t is measured from the beginning of the j^{th} spell. Although spells within the same individual may not be independent, the following partial likelihood is still valid for estimation of the $\boldsymbol{\beta}_j$'s:

$$L(\beta_1, \dots, \beta_K) = \prod_{j=1}^K \prod_{i=1}^{N_j} \left[\frac{e^{\mathbf{x}_{ij}^*(T_{ij})\beta_j}}{\sum_{r=1}^{N_j} Y_{rj}(T_{ij}) e^{\mathbf{x}_{rj}^*(T_{ij})\beta_j}} \right]^{\delta_{ij}}, \quad (11)$$

Here, T_{1j}, \dots, T_{N_jj} are N_j durations of possibly right-censored j^{th} order spells, $\delta_{ij} = 1$ if T_{ij} is an observed duration and $\delta_{ij} = 0$ otherwise, and K is the highest order of spells to be included in the Cox model. The denominator sum is taken over the j^{th} spells that are at risk of being completed at time T_{ij} , i.e., $Y_{rj}(t) = 1$ if $t \leq T_{rj}$, and is equal to 0 otherwise. The corresponding covariate vector observed on $[0, T_{ij}]$ is $\mathbf{x}_{ij}(t)$. Partial likelihood (11) can be maximized separately for each j if there are no additional restrictions on the β_j 's.

The corresponding score equations that define the finite population parameter $\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, \dots, \mathbf{B}'_K)'$ are:

$$U_0(\mathbf{B}) = \sum_{j=1}^K \sum_{i=1}^{N_j} u_{ij0}(T_{ij}, \mathbf{B}_j) = 0, \quad (12)$$

with

$$u_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B}_j)}{S^{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

and with $S^{(0)}(t, \mathbf{B}_j)$ and $S^{(1)}(t, \mathbf{B}_j)$ having the form of (5) and (6) respectively, but with N_j replacing n and \mathbf{B}_j replacing β .

The design-based estimates of the parameters \mathbf{B}_j are obtained by solving equations $\sum_{i=1}^{N_j} w_i(s) \hat{u}_{ij0}(T_{ij}, \hat{\mathbf{B}}_j) = 0$ separately for each j , where \hat{u}_{ij0} has the form of u_{ij0} but with $S^{(0)}$ and $S^{(1)}$ replaced by $\hat{S}^{(0)}$ and $\hat{S}^{(1)}$ respectively. Note that the sampling weights correspond to individuals and not to spells. Similarly, estimation of the covariance matrix of each $\hat{\mathbf{B}}_j$ will be done separately using the design-based robust estimation approach described in Section 2.2. Technically, this is a set of analyses separated by spell order.

Model 2: The second model considered is the marginal model (Wei, Lin and Weissfeld 1989):

$$h_j(t | \mathbf{x}_{ij}(t)) = \lambda_{0j}(t) e^{\mathbf{x}_{ij}^{(t)}\beta_j},$$

where, for each spell order, we allow a different baseline hazard function while the covariate effects are kept the same over different spell orders. The corresponding partial likelihood function as well as the risk set, under the assumption that spells within the same individual are independent, is the same as for Model 1, with β replacing the β_j 's. The corresponding score equation that defines the finite population parameter is

$$U_0^*(\mathbf{B}) = \sum_{j=1}^K \sum_{i=1}^{N_j} u_{ij0}^*(T_{ij}, \mathbf{B}) = 0,$$

with

$$u_{ij0}^*(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S^{(1)}(T_{ij}, \mathbf{B})}{S^{(0)}(T_{ij}, \mathbf{B})} \right\},$$

where $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ are defined by (5) and (6) respectively, but with N_j replacing n and \mathbf{B} replacing β .

The design-based estimate of the parameter \mathbf{B} is obtained by solving the weighted score equations

$$\sum_{i=1}^K \sum_{j=1}^{N_j} w_i(s) \hat{u}_{ij0}^*(T_{ij}, \hat{\mathbf{B}}) = 0,$$

where \hat{u}_{ij0}^* has the form of u_{ij0}^* but with $S^{(0)}(t, \mathbf{B})$ and $S^{(1)}(t, \mathbf{B})$ replaced by $\hat{S}^{(0)}(t, \hat{\mathbf{B}})$ and $\hat{S}^{(1)}(t, \hat{\mathbf{B}})$ respectively.

The estimation of the covariance matrix of $\hat{\mathbf{B}}$ will be done using the design-based robust estimation approach explained in Section 3.2.

Model 3: The last model considered is the following:

$$h_j(t | \mathbf{x}_{ij}) = \lambda_{0j}(t) e^{\mathbf{x}_{ij}^{(t)}\beta_j}.$$

In this model we assume that the baseline hazard functions and the effects of covariates are common for different orders of spells. The risk set at time T_{ij} is defined differently than for Models 1 and 2, and contains all spells with $t \leq T_{ij}$, effectively assuming that all spells are from different individuals. Technically, this model is a single-spell model, so that estimation of coefficients and variances by a design-based robust method is straightforward.

5. Example of modelling multiple unemployment spells

5.1 The data

The data set that we use for illustration comes from the first six-year panel (1993-1998) of the Canadian Survey of Labour and Income Dynamics (SLID). In this panel, about 31,000 individuals from approximately 15,000 households were followed for six years through annual interviews. Some individuals dropped out of the sample over time for any number of reasons while a few others, after missing one or more interviews, resumed their participation. A complex weighting of the responding SLID individuals each year takes into account different types of attrition so that each respondent in a particular year is weighted against the

relevant reference population of 1993. This results in a separate longitudinal weight for each wave (*i.e.*, year) of data. For this analysis we used the longitudinal weights from the last year of the panel, *i.e.*, 1998, which meant that data just from those individuals who were respondents in the final wave of the panel were included in the analyses. A good summary of the sample design issues in SLID is given in Lavigne and Michaud (1998). A review of the issues related to studies of unemployment spells from SLID is given in Roberts and Kovačević (2001).

The state of interest is "being unemployed", defined in this case as the state between a permanent layoff from a full-time job and the commencement of another full-time job. A job is "full-time" if it requires at least 30 hours of work per week. The event of interest is "the exit from unemployment". Only spells beginning after January 1, 1993 were included since January 31, 1993 is the starting date for observations from the panel. Spells that were not completed by the end of the observation period (December 31, 1998) were considered censored. Sample counts of the number of individuals experiencing eligible spells and the number of spells according to their order are given in Table 1. In brief, there were 17,880 spells from 8,401 longitudinal individuals. About half of the sampled individuals (4,260) who became unemployed during this period experienced two or more unemployment spells. There were 3,809 spells that remained uncompleted due to the termination of the panel.

From a long list of available covariates we chose only ten. The variable sex [SEX] of the longitudinal individual is

the only variable that remains constant over different spells. Four variables have values recorded at the end of the year in which the spell commenced: education level [EDUCLEV] with 4 categories (low, low-medium, medium, high), marital status [MARST] with three categories (single, married/common law, other), family income per capita (in Canadian dollars) with 4 categories (<10K, 10K-20K, 20K-30K, 30K+), and age [AGE] (in years). Three variables have the values from the lay-off job preceding the spell: type of job ending [TYPJBEND] with two categories (fired and voluntary), occupation [OCCUPATION] with 6 categories (professional, administration, primary sector, manufacturing, construction, and others); and firm size [FIRMSIZE] with five categories (<20, 20-99, 100-499, 500-999, 1,000+ employees). Two binary variables represent the situation during the spell: having a part time job [PARTTJB], and attending school [ATSCH].

The data set was prepared in the "counting process" style where each individual with eligible spells is represented by a set of rows, and each row corresponds to a spell. Although a row contains time of entry to the spell t_1 , and time of exit t_2 or time of censoring t_c , the duration time for analysis is always considered in the form $(0, t_2 - t_1)$ or $(0, t_c - t_1)$. The covariates under consideration are attached to each row. Also attached to each row are the 1998 longitudinal weight and the identifiers for the stratum and the PSU of the person whose spell is being described by that record.

Table 1 Counts of individuals in the six-year panel of SLID with unemployment spells beginning between January 1993 and December 1998, by the total number of spells and by order of spell (C-completed, U-uncompleted)

Individuals by number of spells	Spells by order									
	First		Second		Third		Fourth		5 th +	
	C	U	C	U	C	U	C	U	C	U
1 spell	4,141	2,221	1,920	-	-	-	-	-	-	-
2 spells	1,915	1,915	-	1,154	761	-	-	-	-	-
3 spells	1,044	1,044	-	1,044	-	612	432	-	-	-
4 spells	629	629	-	629	-	629	-	348	281	-
5+ spells	672	672	-	672	-	672	-	672	-	1,158
Total	8,401	6,481	1,920	3,499	761	1,913	432	1,020	281	1,158
										415

5.2 Analysis

For the purpose of this illustration we restricted the analysis to the first four spells, which means that all sampled individuals with eligible spells are included in the analysis but the spell records after the fourth spell are not considered due to their small number in the sample.

We estimated coefficients and their variances for the 3 models by the design-based methods described in Section 4 through the use of the "SURVIVAL" procedure in SUDAAN Version 8. For all three models, the survey design was specified to be stratified with with-replacement selection of PSU's (*i.e.*, DESIGN = WR). All three models were fit to the same number of spells (16,307). For each model, we then calculated the empirical cumulative baseline hazard functions using a product-limit approach (see Kalbfleisch and Prentice (2002), pages 114-116) as implemented in the SURVIVAL procedure in SUDAAN.

In the robust model-based approach for multiple spells described in Section 3.1, there is an adjustment in the variance estimates to account for the possible dependence among spells from the same individual, assuming independence of spells from different individuals; however, in this approach, no account is made for the unequal probabilities of selection of the sampled individuals - in either the coefficient estimates or the variance estimates. In order to do this, for models 1 and 2 we also used the SURVIVAL procedure in SUDAAN Version 8, to estimate the variances of the weighted coefficient estimates where we assumed independence of spells between individuals but allowed for possible correlation of spells from the same individual. We did this by specifying the sampling design to be unstratified and having with-replacement selection of clusters, and we specified that each individual formed his own cluster. The dependence assumptions are the same as those used by Lin (1994) but we accounted for the use of weights in the estimation of the coefficients and the variances. We will call these variance estimates "modified robust model-based variance estimates of weighted coefficient estimates".

5.3 Some descriptive statistics

The estimated mean duration of a completed spell is 33.3 weeks while the estimated mean duration of the observed portion of a censored (uncompleted) spell is 48.5 weeks.

Visual examination of estimated Kaplan-Meier survival functions (not shown) for spells of each order indicated that, as order increased, the value of the survivor function at any fixed time t decreased, indicating that first spells are the

longest among completed spells, and that the higher the order of a multiple spell the shorter is its duration. This is likely to be a consequence of the limited life of the panel, in the sense that an individual with more spells in the given six-year time frame is likely to have shorter spells.

5.4 Model fits using a design-based approach

As noted earlier, our example is just an illustration of the design-based approach to fitting proportional hazards models to multiple-event data from a survey with a complex design. Thus, little time is spent in this article on discussing how to assess the adequacy of these models, such as the adequacy of the proportionality assumptions in all of the models or whether one type of model fits as well as another.

Estimated coefficients from fitting the three models to the SLID data are given in Table 2. Coefficients found significant at the 5% level, through the use of individual t tests, are shown in bold.

Model 1 is conditional on the spell order and involved fitting four models separately to the data from the four different spell orders. As seen in Table 2, SEX, AGE, and at least one category of the Family Income variable were significant for spells of all orders, although magnitudes of the estimated coefficients differed with the spell order. The estimated coefficients for AGE were negative but decreased in magnitude as the spell order increased, while there was no discernable pattern in the estimated coefficients for the other 2 variables. The variables EDUCLEV, PARTJB and ATSCH had significant coefficients for spells of order 1, 2, and 3, but not for spells of order 4. This can be at least partly attributed to the small sample size for the fourth spells. For each of the other three variables in the model (MARST, OCCUPATION, and FIRMSIZE), there was just one spell order for which a coefficient was significant.

For Model 2, the model coefficients are restricted to be the same for all spell orders. As seen in Table 2, numerically many - but not all - of the estimated coefficient values were situated between the estimates for the first and the second spells obtained for Model 1 which could be due to the fact that a high proportion of the sample corresponded to events of these orders. All but the OCCUPATION variable had a significant coefficient. Standard errors of coefficients were smaller for Model 2 than for Model 1.

Model 3 is a single-spell model with a single set of model coefficients and a single baseline hazard function. The estimated model coefficients are similar to the estimates obtained by Model 2.

Table 2 Estimated β coefficients for three models

	Model 1				Model 2	Model 3
	Order 1	Order 2	Order 3	Order 4		
SEX (F)						
M	0.4417	0.3781	0.3299	0.4435	0.4049	0.4090
EDUCLEV (H)						
L	-0.4561	-0.5234	-0.3748	-0.1065	-0.4128	-0.4331
LM	-0.2330	-0.2700	-0.3310	-0.1653	-0.2436	-0.2474
M	-0.0744	-0.1060	-0.1156	0.0668	-0.0684	-0.0671
MARST (M)						
Single	-0.1142	-0.1290	-0.0622	-0.1375	-0.1357	-0.1330
Other	0.0985	-0.0894	0.1124	-0.1072	0.0328	0.0401
TYPJBEND (Fired)						
Voluntary	0.0704	0.2752	0.4207	0.3413	0.1579	0.1284
OCCUPATION(Othrs)						
Professionals	0.1592	-0.1364	-0.1388	0.0903	0.0490	0.0485
Admin	-0.0265	-0.2930	-0.1769	0.0579	-0.0971	-0.0938
PrimSector	-0.0211	-0.2175	-0.1187	0.2032	-0.0410	-0.0201
Manufacture	-0.0003	-0.0994	-0.1295	0.2862	-0.0093	-0.0088
Construction	0.1290	-0.1862	-0.0879	0.2339	0.0490	0.0813
FIRMSIZE (1000+)						
<20	-0.0027	-0.0097	0.1005	0.4403	0.0441	0.0408
20-99	0.0358	0.0881	0.0815	0.3999	0.0928	0.0951
100-499	0.0436	-0.0905	0.0328	0.0257	0.0214	0.0278
500-999	-0.0006	0.0153	-0.0623	-0.0067	-0.0005	0.0020
PARTTJB (No)						
Yes	-0.2903	-0.5414	-0.5109	-0.1407	-0.3693	-0.3743
ATSCH (No)						
Yes	-1.0832	-1.1516	-1.2956	-1.3541	-1.1205	-1.1266
Family Income Per Capita (10K-)						
10K-20K	0.1294	0.1802	0.0692	0.1117	0.1345	0.1330
20K-30K	0.1644	0.3611	0.1572	0.4900	0.2241	0.2141
30K+	0.1712	0.3916	0.3005	0.4241	0.2280	0.2115
AGE	-0.0491	-0.0311	-0.0269	-0.0207	-0.0424	-0.0435
Spells in risk set	8,386	4,255	2,345	1,300	16,286	16,286
Censored	1,913	759	432	281	3,385	3,385
Completed	6,473	3,496	1,913	1,019	12,901	12,901

The values significant at a 5% level are bold.

The estimated cumulative baseline hazard functions for Models 1 to 3 are given in Figures 1 to 3 respectively. In all cases, for durations up to approximately 50 weeks, the functions have a concave shape, implying that there is a positive time dependence of the exit rate (*i.e.* the longer the spell, the higher the probability of exit). For durations longer than 50 weeks, the shapes become convex, suggesting negative time dependence for the longer spells. In Figure 1, positions of the estimated cumulative baseline hazard functions vary according to spell order, with the curve for spells of order 1 being the highest, and the curve for spells of order 4 being the lowest. In Figure 2, for Model 2, the positions of the different curves do not follow spell order. This observed difference between Figures 1 and 2 could serve as one visual diagnostic that further study is required in order to assess whether Model 1 or Model 2 is a better descriptor of the data, since estimated coefficients have an impact on the estimated baseline hazards.

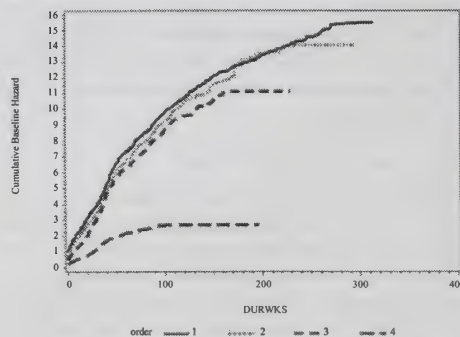


Figure 1 Cumulative Baseline Hazard – Model 1

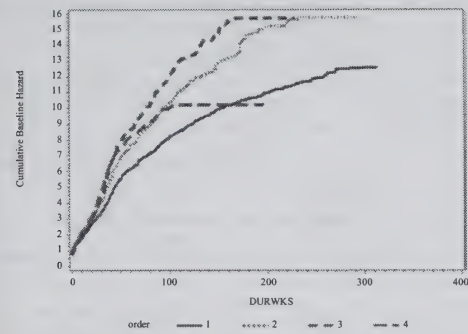


Figure 2 Cumulative Baseline Hazard – Model 2

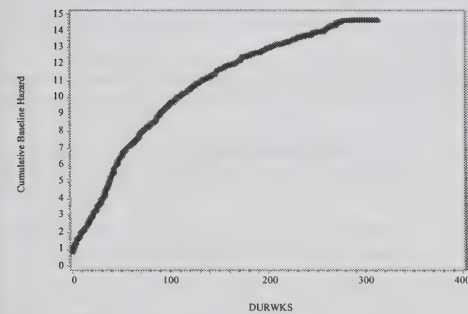


Figure 3 Cumulative Baseline Hazard – Model 3

5.5 Comparison to modified robust model-based variance estimates

As described in Section 5.2, the modified robust model-based variance estimates account for possible correlation among spells from the same individual, where independence among individuals is assumed. When, for Models 1 and 2, the standard error estimates obtained by this approach were compared to the design-based standard error estimates, only very minor differences were observed. This would seem to indicate that the design-based estimates are picking up any correlation among spells from the same individual and also that there does not appear to be additional dependence above the level of the individual for our particular example.

6. Concluding remarks

We explored the problem of analysis of multiple spells by considering two general approaches for dealing with the lack of independence among the exit times: a robust model-based approach and a design-based approach. The first approach estimates the model parameters assuming independence of the spells, and then corrects the naïve covariance matrix to account for within-individual dependencies postulated by the researcher. This approach does not

account for the possible clustering between individuals (or, in fact, for any clustering that might occur at a level above the individual) nor for the unequal probabilities of selection of individuals (although, in our example, we showed how the method could be extended to include the survey weights). The second approach defines the model coefficients as finite population parameters. These parameters are then estimated accounting for possible unequal selection probabilities of individuals. A design-based variance estimation method that accounts for possible correlations between individuals in the same PSU automatically accounts for the unspecified dependencies of spells at levels below the PSU, such as dependencies within individuals. For large sample sizes this design-based inference extends directly to the super-population from which, hypothetically, the finite population was generated. The deficiency of the first approach is that it totally ignores the potential for clustering between individuals. A possible disadvantage of the second approach, as we applied it, is that it relies on the assumption of with-replacement sampling of PSU's of individuals. The two approaches coincide in the case of simple random sampling of individuals where, in the robust model-based approach, dependence among spells from the same individual is explicitly postulated and accounted for in the variance estimation formula and where, in the design-based approach, spells from the same individual are treated as a cluster in the design-based variance estimation.

We applied the design-based approach to three proportional-hazards-type models. One model allowed for differential unspecified baseline hazards and different coefficients for each spell order. The second model still allowed for differential unspecified baseline hazards for different spell orders but required the coefficients to be the same over orders. The third model was a simple single-spell model. We found that how information on the spell order was used affected the results of our model-fitting. A visual comparison of the coefficient estimates and the estimates of the cumulative baseline hazards for Models 1 and 2 indicated different results. A formal test for whether the coefficients actually differ by spell order (as allowed in Model 1), given baseline hazards that can differ by spell order, would be useful, as suggested by one of the referees. It is actually straightforward to produce such a test, and can be done as follows. Let $\gamma = (B'_1, B'_2, \dots, B'_K)'$ be the vector of all K coefficient vectors of Model 1, where each has length p , and let $z_{ij}(t) = (0', 0', \dots, x_{ij}(t)', 0', \dots, 0')'$ be the vector of length pK for the j^{th} spell of the i^{th} individual where the j^{th} component of this vector contains the vector of covariates $x_{ij}(t)$. Then, Model 1 can be expressed as

$$h_j(t | z_{ij}(t)) = \lambda_{0j}(t) e^{z_{ij}(t)\gamma},$$

which has the general form of baseline hazards varying with spell order but a fixed coefficient vector. A test for constancy of the coefficients pertaining to each spell order, i.e., $H_0: \mathbf{B}_1 = \mathbf{B}_2 = \dots \mathbf{B}_K$ is equivalent to testing $H_0: \mathbf{C}\boldsymbol{\gamma} = \mathbf{0}$ where \mathbf{C} is the $(K-1)p \times Kp$ matrix $\mathbf{C} = \mathbf{I}_p \otimes [\mathbf{1}_{K-1} \ -\mathbf{I}_{K-1}]$. Given an estimate $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ and an estimate $\hat{V}(\hat{\boldsymbol{\gamma}})$ of the covariance matrix of $\hat{\boldsymbol{\gamma}}$, obtained as described in Section 4 for Model 2, a Wald statistic may be calculated in order to test the hypothesis. If the hypothesis is not rejected, it may be concluded that a model with constant coefficients over spell order (but baseline hazard varying with spell order) appears to fit the data as well as a model where both baseline hazard and coefficients vary with spell order. Other measures for model adequacy should also be straightforward to develop under the design-based framework.

We also visually compared, for our example, coefficient standard error estimates obtained under the design-based approach (accounting for clustering at the PSU level and lower) and obtained under a modification of the robust model-based approach (accounting for clustering at the individual level and lower) for Models 1 and 2. We found only minor differences, which indicated no clustering effects above the individual level for these particular data. We also calculated standard error estimates assuming independence even between spells from the same person and again found only minor differences with those obtained from the design-based approach. It thus seems that, for this particular example, there is little inter-spell dependence. However, in general, we feel that a design-based approach guards against missing any unpostulated dependencies at the PSU level and lower in the variance estimates.

Acknowledgements

We are grateful to Normand Laniel and Xuelin Zhang for their useful comments to an earlier version of this manuscript. We also thank the associate editor and the referees for comments and suggestions improving greatly the readability of the manuscript.

References

- Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-291.
- Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79, 139-147.
- Blossfeld, H.-P., and Hamerle, A. (1989). Using Cox models to study multipisode processes. *Sociological Methods and Research*, 17, 4, 432-448.
- Clayton, D., and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *Journal of the Royal Statistical Society, Series A*, 1985, 148, 82-117.
- Cochran, W.G. (1977). *Sampling Techniques*. Third edition. New York: John Wiley & Sons, Inc.
- Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.
- Hamerle, A. (1989). Multiple-spell regression models for duration data. *Applied Statistics*, 38, 1, 127-138.
- Heckman, J., and Singer, B. (1982). Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*, (Eds., K. Land and A. Rogers), New York: Academic Press, 567-599.
- Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, 55, 1, 13-22.
- Kalbfleisch, J.D., and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Edition, New York: John Wiley & Sons, Inc.
- Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.
- Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.
- Lavigne, M., and Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Working Paper, Statistics Canada, 75F0002M No. 98-05.
- Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.
- Lin, D.Y. (2000). On Fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.
- Lin, D.Y., and Wei, L.J. (1989). The robust Inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.
- Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.
- Roberts, G., and Kovačević, M. (2001). New research problems in analysis of duration data arising from complexities of longitudinal surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 111-116.
- Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.
- Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.

Bayesian weight trimming for generalized linear regression models

Michael R. Elliott¹

Abstract

In sample surveys where units have unequal probabilities of inclusion in the sample, associations between the probability of inclusion and the statistic of interest can induce bias. Weights equal to the inverse of the probability of inclusion are often used to counteract this bias. Highly disproportional sample designs have large weights, which can introduce undesirable variability in statistics such as the population mean estimator or population regression estimator. Weight trimming reduces large weights to a fixed cutpoint value and adjusts weights below this value to maintain the untrimmed weight sum, reducing variability at the cost of introducing some bias. Most standard approaches are ad-hoc in that they do not use the data to optimize bias-variance tradeoffs. Approaches described in the literature that are data-driven are a little more efficient than fully-weighted estimators. This paper develops Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. An application to estimate injury risk of children rear-seated in compact extended-cab pickup trucks using the Partners for Child Passenger Safety surveillance survey is considered.

Key Words: Sample survey; Sampling weights; Weight Winsorization; Bayesian population inference; Weight smoothing; Generalized linear mixed models.

1. Introduction

Analysis of data from samples with differential probabilities of inclusion typically use case weights equal to the inverse of the probability of inclusion to reduce or remove bias in the estimators of population quantities of interest. Replacing implicit means and totals in statistics with their case-weighted equivalents yields unbiased linear estimators and asymptotically unbiased non-linear estimators of population values (Binder 1983). Case weights may also incorporate non-response adjustments, which typically are equal to the inverse of the estimated probability of response (Gelman and Carlin 2002, Oh and Scheuren 1983), or calibration adjustments, which constrain case weights to equal known population totals, either jointly, as in poststratification or generalized regression estimation, or marginally, as in generalized raking estimation (Deville and Särndal 1992, Isaki and Fuller 1982).

There is little debate that sampling weights be utilized when considering descriptive statistics such as means and totals obtained from unequal probability-of-selection designs. However, when estimating "analytical" quantities (Cochran 1977, page 4) that focus on associations between, e.g., risk factors and health outcomes estimated via linear and generalized linear models, the decision to use sampling weights is less definitive (*cf* Korn and Graubard 1999, pages 180-182). In a regression setting, discrepancies between weighted and unweighted regression slope estimators can occur either because the data model is misspecified or there is an association between the residual errors and/or the probability of inclusion (sampling is

informative). When the data model is misspecified, one option is to improve the model specification. However, it may be difficult to determine the exact functional form; or it may be that the degree of misspecification is very modest but is magnified by the sample design; or it may be that an approximation to the true model is desired to simplify explanation (linearly approximating a quadratic trend). In the case of informative or non-ignorable sampling, design weights may be required to obtain consistent estimators of regression parameters (Korn and Graubard 1995). More formally, fully-weighted estimators of regression parameters are "pseudo-maximum likelihood" estimators (PMLEs) (Binder 1983, Pfeffermann 1993) in that they are "design consistent" for MLEs that would solve the score equations for the regression parameters under the assumed superpopulation regression model if we had observed data for the entire population. Design consistency implies that the difference between the population target quantity and the estimate derived from the sample tends to zero as the sample size and population size jointly increase, or that these differences will on average tend to 0 from repeated sampling of the population, where samples are selected in an identical fashion from $t \rightarrow \infty$ replicates of the population: see Särndal (1980) or Isaki and Fuller (1982). If observations are clustered, more care must be taken to develop design consistent estimators of PLMEs, although nested multi-stage designs allow for the census log-likelihood estimates to be approximated using weighted score equations if care is taken to account for the fact that the within-cluster sample sizes typically are small and remain so even if the number of clusters increases

1. Michael R. Elliott is Assistant Professor, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI. E-mail: mre Elliott@umich.edu.

(Pfeffermann, Skinner, Holmes, Goldstein and Rabash 1998, Korn and Graubard 2003).

Although PMLEs are popular because of design consistency, this property is purchased at the cost of increased variance. This increase can overwhelm the reduction in bias, so that the MSE actually increases under a weighted analysis. This is particularly likely if a) the sample size is small, b) the differences in the inclusion probabilities are large, or c) the model is approximately correctly specified and the sampling is approximately noninformative. Perhaps the most common approach to dealing with this problem is *weight trimming* (Potter 1990, Kish 1992, Alexander, Dahl and Weidman 1997), in which weights larger than some value w_0 are fixed as w_0 . Typically w_0 is chosen in an *ad hoc* manner - say 3 or 6 times the mean weight - without regard to whether the chosen cutpoint is optimal with respect to MSE. Thus bias is introduced to reduce variance, with the goal of an overall reduction in MSE.

Other design-based methods have been considered in the literature. Potter (1990) discusses systematic methods for choosing w_0 , including weight distribution and MSE trimming procedures. The weight distribution technique assumes that the weights follow an inverted and scaled beta distribution; the parameters of the inverse-beta distribution are estimated by method-of-moment estimators, and weights from the upper tail of the distribution, say where $1 - F(w_i) < 0.01$, are trimmed to w_0 such that $1 - F(w_0) = 0.01$. The MSE trimming procedure determines the empirical MSE at trimming level w_i , where the trimmed weight $w_i^* = w_i I(w_i \geq w_i) + w_0 I(w_i < w_i)$, $i = 1, \dots, n$ under the assumption that the fully weighted estimate is unbiased for the true mean. In practice, one considers a variety of trimming levels $t = 1, \dots, T$, where $t = 1$ corresponds to the unweighted data ($w_i = \min_i(w_i)$) and $t = T$ to the fully-weighted data ($w_T = \max_i(w_i)$), and $\hat{\theta}_t$ is the value of the statistic using the trimmed weights at level t . The trimming level chosen is then given by $w_0 = w_{t^*}$, where $t^* = \operatorname{argmin}_t(\operatorname{MSE}_t)$ for $\operatorname{MSE}_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\hat{\theta}_t)$.

In the calibration literature, techniques have been developed that allow generalized poststratification or raking adjustments to be bounded to prevent the construction of extreme weights (Deville and Särndal 1992, Folsom and Singh 2000). Beaumont and Alavi (2004) extend this idea to develop estimators that focus on trimming large weights of highly influential or outlying observations. While these bounds trim extreme weights to a fixed cutpoint value, the choice of this cutpoint remains arbitrary.

An alternative approach to the direct weight trimming procedures has been developed in the Bayesian finite population inference literature (Elliott and Little 2000, Holt

and Smith 1979, Ghosh and Meeden 1986, Little 1991, 1993, Lazzeroni and Little 1998, Rizzo 1992). These approaches account for unequal probabilities of inclusion by considering the case weights as stratifying variables within strata defined by the probability of inclusion. These "inclusion strata" may correspond to formal strata from a disproportional stratified sample design, or may be "pseudo-strata" based on collapsed or pooled weights derived from selection, poststratification, and/or non-response adjustments. Standard weighted estimates are then obtained when the weight stratum means of survey outcomes are treated as fixed effects, and trimming of the weights is achieved by treating the underlying weight stratum means as random effects. These methods allow for the possibility of "partially-weighted" data that uses the data itself to appropriately modulate the bias-variance tradeoff, and also allows estimation and inference from data collected under unequal probability-of-inclusion sample designs to be based on models common to other fields of statistical estimation and inference.

This paper extends these random-effects models, which we term "weight smoothing" models, to include estimation of population parameters in linear and generalized linear models. Section 2 briefly reviews Bayesian finite population inference, formalizes the concept of ignorable and non-ignorable sampling mechanisms, and develops the weight smoothing models for linear and generalized linear regression models in a fully Bayesian setting. Section 3 provides simulation results to consider the repeated sampling properties of the weight smoothing estimators of linear and logistic regression parameters in a disproportional-stratified sample design and compares them with standard design-based estimators. Section 4 illustrates the use of the weight smoothing estimators in an analysis of risk of injury to children in passenger vehicle crashes. Section 5 summarizes the results of the simulations and considers extensions to more complex sample designs.

2. Bayesian finite population inference

Let the population data for a population with $i = 1, \dots, N$ units be given by $Y = (y_1, \dots, y_N)$, with associated covariate vectors $X = (x_1, \dots, x_N)$ and sampling indicator variable $I = (I_1, \dots, I_N)$, where $I_i = 1$ if the i^{th} element is sampled and 0 otherwise. As in design-based population inference, Bayesian population inference focuses on population quantities of interest $Q(Y)$, such as population means $Q(Y) = \bar{Y}$ or population least-squares regression parameters $Q(Y, X) = \min_{B_0, B_1} \sum_{i=1}^N (y_i - B_0 - B_1 x_i)^2$. In contrast to design-based inference, but consistent with most other areas of statistics, one posits a model for the population data Y as a function of parameters θ :

$Y \sim f(Y|\theta)$. Inference about $Q(Y)$ is made based on the posterior predictive distribution of $p(Y_{\text{nob}} | Y_{\text{obs}}, I)$, where Y_{nob} consists of the elements of Y_i for which $I_i = 0$:

$$p(Y_{\text{nob}} | Y_{\text{obs}}, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}} \quad (1)$$

where $p(I | Y, \theta, \phi)$ models the inclusion indicator. If we assume that ϕ and θ are *a priori* independent and if the distribution of sampling indicator I is independent of Y , the sampling design is said to be “unconfounded” or “noninformative”; if the distribution of I depends only on Y_{obs} , then the sampling mechanism is said to be “ignorable” (Rubin 1987), equivalent to the standard missing data terminology (the unobserved elements of the population can be thought of as missing by design). Under ignorable sampling designs, $p(\theta, \phi) = p(\theta)p(\phi)$ and $p(I | Y, \theta, \phi) = p(I | Y_{\text{obs}}, \phi)$, and thus (1) reduces to

$$\frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta dY_{\text{nob}}} = p(Y_{\text{nob}} | Y_{\text{obs}}), \quad (2)$$

allowing inference about $Q(Y)$ to be made without explicitly modeling the sampling inclusion parameter I (Ericson 1969, Holt and Smith 1979, Little 1993, Rubin 1987, Skinner, Holt and Smith 1989). Noninformative sample designs are a special case of ignorable sample designs, equivalent to missing completely at random mechanisms being a special case of missing at random mechanisms.

In the regression setting, where inference is desired about parameters that govern the distribution of Y conditional on fixed and known covariates X , (1) becomes

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}}$$

which reduces to

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X) = \frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta dY_{\text{nob}}}$$

if and only if I depends only on (Y_{obs}, X) , of which dependence on X only is a special case. Thus if inference is desired about a regression parameter $Q(Y, X)$, then a noninformative or more generally ignorable sample design

can allow inclusion probabilities to be a function of the fixed covariates.

2.1 Accommodating unequal probabilities of inclusion

Maintaining the ignorability assumption for the sampling mechanism often requires accounting for the sample design in both the likelihood and prior model structure. In the case of the unequal probability-of-inclusion sample designs, this can be accomplished by developing an index $h = 1, \dots, H$ of the probability of inclusion (Little 1983, 1991); this could either be a one-to-one mapping of the case weight order statistics to their rankings, or a preliminary “pooling” of the case weights using, e.g., the $100/H$ percentiles of the case weights. The data are then modeled by

$$y_{hi} | \theta_h \sim f(y_{hi} ; \theta_h), \quad i = 1, \dots, N_h$$

for all elements in the h^{th} inclusion stratum, where θ_h allows for an interaction between the model parameter(s) θ and the inclusion stratum h . Putting a noninformative prior distribution on θ_h then reproduces a fully-weighted analysis with respect to the expectation of the posterior predictive distribution of $Q(Y)$.

To make this concrete, assume we are interested in estimating a population mean $Q(Y) = \bar{Y} = N^{-1} \sum_{i=1}^N y_i$ from a unequal probability-of-inclusion sample with a simple random sample within inclusion strata. Rewriting as $Q(Y) = \sum_h P_h \bar{Y}_h$ where $\bar{Y}_h = N_h^{-1} \sum_{i=1}^{N_h} y_{hi}$ is the population inclusion stratum mean and $P_h = N_h/N$, we have

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h \{ n_h \bar{y}_{h, \text{obs}} + (N_h - n_h) E(\bar{Y}_h | Y_{\text{obs}}) \}$$

where \bar{Y}_h is decomposed into the observed inclusion stratum mean $\bar{y}_{h, \text{obs}} = n_h^{-1} \sum_{i=1}^{N_h} I_{hi} y_{hi}$ and the unobserved inclusion stratum mean $\bar{Y}_{h, \text{nob}} = (N_h - n_h)^{-1} \sum_{i=1}^{N_h} (1 - I_{hi}) y_{hi}$. If we assume

$$y_{hi} | \mu_h, \sigma_h^2 \stackrel{\text{ind}}{\sim} N(\mu_h, \sigma_h^2)$$

$$p(\mu_h, \sigma_h^2) \propto 1$$

then

$$E(\bar{Y}_{h, \text{nob}} | Y_{\text{obs}}) = E(E(\bar{Y}_{h, \text{nob}} | Y_{\text{obs}}) | Y_{\text{obs}}, \mu_h, \sigma_h^2) = E(\mu_h | Y_{\text{obs}}) = \bar{y}_{h, \text{obs}}.$$

and the posterior predictive mean of the population mean is given by the weighted sample mean:

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h N_h \bar{y}_{h, \text{obs}} = N^{-1} \sum_{i=1}^N I_{hi} w_{hi} y_{hi}$$

where $w_{hi} \equiv w_h = N_h/n_h$ for all the observed elements in inclusion stratum h . Further, the weighted mean will be the posterior predictive expectation of the population mean for any assumed distribution of Y as long as $E(y_{hi} | \mu_h) = \mu_h$. In contrast, a simple exchangeable model for the data

$$y_i | \mu, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mu, \sigma^2)$$

$$p(\mu, \sigma^2) \propto 1$$

yields $E(\bar{Y} | Y_{\text{obs}}) = n^{-1} \sum_{i=1}^N I_i y_i$, the unweighted estimator of the mean, which may be badly biased if exchangeability fails to hold, as would be the case if there is an association between the probability of inclusion and Y .

2.2 Weight smoothing models

In its general form, our proposed “weight smoothing method” stratifies the data by the probability of inclusion and then uses a hierarchical model to effect trimming via shrinkage. A general description of such a model is given by

$$y_{hi} | \theta_h \sim f(y_{hi}; \theta_h) \quad (3)$$

$$\theta_h | M_h, \mu, R \sim N(\hat{y}_h, R), \hat{y}_h = g(M_h, \mu)$$

$$\mu, R | M_h \sim \Pi.$$

where $h = 1, \dots, H$ indexes the probability of inclusion from the highest to the lowest probabilities, $g(M_h, \mu)$ is a function linking information M_h from the inclusion probability stratum and a smoothing parameter μ to the data distribution parameter θ_h indexed by the inclusion stratum, and Π is a flat or weakly informative hyperparameter distribution (Little 2004).

The particulars of the likelihood and prior specifications will depend on the population parameter of interest, the sample design, distributional assumptions about y , and efficiency-robustness tradeoffs. Positing an exchangeable model on the inclusion stratum means from the previous example yields (Lazzeroni and Little 1998, Elliott and Little 2000)

$$y_{hi} | \theta_h \stackrel{\text{ind}}{\sim} N(\theta_h, \sigma^2)$$

$$\theta_h \stackrel{\text{ind}}{\sim} N(\mu, \tau^2).$$

Assuming for the moment σ^2 and τ^2 known, we have

$$E(\bar{Y} | Y_{\text{obs}}) = N^{-1} \sum_h \{ n_h \bar{y}_{h, \text{obs}} + (N_h - n_h) E(\mu_h | Y_{\text{obs}}) \}$$

where $E(\mu_h | Y_{\text{obs}}) = w_h \bar{y}_h + (1 - w_h) \bar{y}$ for $w_h = \tau^2 n_h / (\tau^2 n_h + \sigma^2)$ and $\bar{y} = (\sum_h n_h / (n_h \tau^2 + \sigma^2))^{-1} \sum_h n_h / (n_h \tau^2 + \sigma^2) \bar{y}_h$. As $\tau^2 \rightarrow \infty$, $w_h \rightarrow 1$ so that $E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h \bar{y}_h$; thus a flat prior recovers the fully-weighted estimator, as we showed previously. On the other hand, as $\tau^2 \rightarrow 0$, $w_h \rightarrow 0$ so that $E(\mu_h | Y_{\text{obs}}) \rightarrow \bar{y} |_{\tau^2=0} = \bar{y}$, the unweighted mean; thus the excluded units of the sample are estimated at the pooled mean since the model assumes that all y_{hi} are drawn from a common mean. Hence this weight smoothing model allows compromise between the design-consistent estimator which may be highly inefficient, and the unweighted estimator that is fully efficient under the strong assumption that the inclusion probability and mean of Y are independent. By assuming a weak hyperprior distribution on τ^2 , the degree of compromise between the weighted and unweighted mean will be “data-driven,” albeit under the modeling assumptions.

2.3 Weight smoothing for linear and generalized linear regression models

Generalized linear regression models (McCullagh and Nelder 1989) postulate a likelihood for y_i of the form

$$f(y_i; \theta_i, \phi) = \exp \left[\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right] \quad (4)$$

where $a_i(\phi)$ involves a known constant and a (nuisance) scale parameter ϕ , and the mean of y_i is related to a linear combination of fixed covariates \mathbf{x}_i through a link function $g(\cdot)$: $E(y_i | \theta_i) = \mu_i$, where $g(\mu_i) = g(b'(\theta_i)) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$. We also have $\text{Var}(y_i | \theta_i) = a_i(\phi) V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$. The link is canonical if $\theta_i = \eta_i$, in which case $g'(\mu_i) = V^{-1}(\mu_i)$. Well-known examples are the normal distribution, where $a_i(\phi) = \sigma^2$ and the canonical link is $g(\mu_i) = \mu_i$; the binomial distribution, where $a_i(\phi) = n_i^{-1}$ and the canonical link is $g(\mu_i) = \log(\mu_i / (1 - \mu_i))$; and the Poisson distribution, where $a_i(\phi) = 1$ and the canonical link is $g(\mu_i) = \log(\mu_i)$.

Indexing the inclusion stratum by h , we have $g(E[y_{hi} | \boldsymbol{\beta}_h]) = \mathbf{x}_{hi}^T \boldsymbol{\beta}_h$. We assume a hierarchical model of the form

$$(\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_H^T)^T | \boldsymbol{\beta}^*, G \sim N_{Hp}(\boldsymbol{\beta}^*, G). \quad (5)$$

where $\boldsymbol{\beta}^*$ is an unknown vector of mean values for the regression coefficients and G is an unknown covariance matrix.

We consider the target population quantity of interest $\mathbf{B} = (B_1, \dots, B_p)^T$ to be the slope that solves the population score equation $U_N(\mathbf{B}) = 0$ where

$$U_N(\beta) = \sum_{i=1}^N \frac{\partial}{\partial \beta} \log f(y_i; \beta) = \sum_{h=1}^H \sum_{i=1}^{N_h} \frac{(y_{hi} - g^{-1}(\mu_i(\beta)))x_{hi}}{V(\mu_{hi}(\beta))g'(\mu_{hi}(\beta))}. \quad (6)$$

Note that the quantity B such that $U(B) = 0$ is always a meaningful population quantity of interest even if the model is misspecified (i.e., η_i is not exactly linear with respect to the covariates), since it is the linear approximation of x_i to $\eta_i = g(\mu_i)$. Under the model given by (4) and (5), a first-order approximation (assuming a negligible sampling fraction) to $E(B | y, X)$ is given by \hat{B} where

$$\sum_{h=1}^H W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{B})))x_{hi}}{V(\mu_{hi}(\hat{B}))g'(\mu_{hi}(\hat{B}))} = 0 \quad (7)$$

where $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}(x_{hi}^T \hat{\beta}_h)$, and $\hat{\beta}_h = E(\beta_h | y, X)$, as determined by the form of (5). (If N_h is unknown, it can be replaced with $\hat{N}_h = \sum_{i \in h} w_{hi}$, and the $\hat{N}_1, \dots, \hat{N}_H$ treated as a multinomial distribution of size N parameterized by unknown inclusion stratum probabilities q_1, \dots, q_H with, e.g., a Dirichlet prior.) Thus, in the example of linear regression, where $V(\mu_i) = \sigma^2$ and $g'(\mu_i) = 1$, (7) resolves to

$$\hat{B} = E(B | y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} x_{hi} x_{hi}^T \right]^{-1} \left[\sum_h W_h \left(\sum_{i=1}^{n_h} x_{hi} x_{hi}^T \right) \hat{\beta}_h \right]. \quad (8)$$

In the example of logistic regression, where $V(\mu_i) = \mu_i(1 - \mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1 - \mu_i)^{-1}$, $E(B | y, X)$ is given by solving for the population regression parameters β_j , $j = 1, \dots, p$

$$\sum_{h=1}^H W_h \sum_{i=1}^{n_h} x_{hi} \frac{\exp(x_{hi} \beta_j)}{1 + \exp(x_{hi} \beta_j)} = \sum_{h=1}^H W_h \sum_{i=1}^{n_h} x_{hi} \frac{\exp(x_{hi} \hat{\beta}_j)}{1 + \exp(x_{hi} \hat{\beta}_j)}. \quad (9)$$

This can be accomplished via simple root-finding numerical methods such as Newton's Method.

We consider four forms of β^* and G in (5) in this paper:

1. Exchangeable Random Slope (XRS):
 $\beta_h^* = (\beta_0^*, \dots, \beta_p^*)$ for all h , $G = I_H \otimes \Sigma$. (10)
2. Autoregressive Random Slope (ARS):
 $\beta_h^* = (\beta_0^*, \dots, \beta_p^*)$ for all h ,
 $G = A \otimes \Sigma$, $A_{jk} = \rho^{|j-k|}$, $j, k = 1, \dots, H$.
3. Linear Random Slope (LRS):
 $\beta_h^* = (\beta_{00}^* + \beta_{01}^* h, \dots, \beta_{p0}^* + \beta_{p1}^* h)$,
 $G = I_H \otimes \Lambda$.

4. Nonparametric Random Slope (NPRS):

$$\beta_h^* = (f_0(h), \dots, f_p(h)), G = 0.$$

$$\left\{ \begin{array}{l} f_j : f_j^v \text{ absolutely continuous, } v = 0, 1, \\ \int (f_j^{(2)}(u))^2 du < \infty, \\ \min_j \sum_h (\beta_{hj}^* - f_j(h))^2 + \lambda_j \int (f_j^{(2)}(u))^2 du \end{array} \right\}$$

where h again indexes the probability of inclusion, I_H is an $H \times H$ identity matrix, ρ is an autocorrelation parameter that controls the degree of shrinkage across the weight strata, Σ is an unconstrained $p \times p$ covariance matrix, Λ is a $p \times p$ diagonal matrix, and $f_j(h)$ is a twice differentiable smooth function of h that minimizes the residual sum of squares plus a roughness penalty parameterized by λ_j (Wahba 1978, Hastie and Tibshirani 1990). Reformulating the NPRS model as in Wang (1998) we have

$$y_{hi} | \beta_h^* \sim N(x_{hi}^T \beta_h^*, \sigma^2)$$

$$\beta_{hj} = \beta_{j0}^* + \beta_{j1}^* h + \omega_{hj}$$

$$u_j \sim N_{H-1}(0, I \tau_j^2), \tau_j^2 = \sigma^2 / (H \lambda_j) \quad j = 0, \dots, p$$

where ω_h is the h^{th} row of Choleski decomposition of the cubic spline basis matrix Ω where $\Omega_{hk} = \int_0^1 ((h-1)/(H-1-t)_+)((k-1)/(H-1-t)_+ dt)$, $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$, $h, k = 1, \dots, H$. The NPRS model can be extended into the generalized linear model form as in Lin and Zhang (1999), where the first-stage normality assumption is replaced with a link function that is linear in the covariates: $g(E(y_{hi} | \beta_h)) = x_{hi}^T \beta_h$, for $g(\cdot)$ as in (4).

Assuming for the moment that the second stage parameters are known, we see that, in the case of the XRS model with normal data, as $|G| \rightarrow \infty$, sharing of information across inclusion strata ceases, and $\hat{\beta}_h \approx (x_h^T x_h)^{-1} x_h^T y_h$, the regression estimator within the inclusion stratum. Replacing this into (8) yields $\hat{B} \approx \hat{B}^*$, the fully weighted estimator of the population slope. Similarly, as $|G| \rightarrow 0$, the within-inclusion-stratum slopes $\hat{\beta}_h \approx \beta^*$ the common prior slope, yielding $\hat{B} \approx \beta^*$ when replaced in (8), or \hat{B}^* if a non-informative hyperprior distribution is placed on β^* and its posterior mean obtained as $(x^T x)^{-1} x^T y$. Empirical or fully Bayesian methods that allow the data to estimate the second stage parameters thus allow for data-driven "weight smoothing," compromising between the unweighted and fully-weighted estimators.

In practice, of course, the second-stage mean and variance components are usually not known; hence we

complete the model specification by postulating a hyperprior distribution for the second-stage parameters:

$$p(\phi, \beta^*, G) \propto p(\zeta).$$

Typically the hyperprior distribution $p(\zeta)$ is either weakly informative or non-informative. Gibbs sampling (Gelfand and Smith 1990; Gelman and Rubin 1992) can then be utilized to obtain draws from the full joint posterior of $(\beta, \beta^*, \phi, G)^T | y, X$. In the XRS model, we consider $p(\sigma, \beta^*, \Sigma) \propto \sigma^{-2} | \Sigma |^{-(p+1/2)} \exp(-1/2 \text{tr}\{\sigma^{-2} \Sigma^{-1}\})$, that is, non-informative prior distributions on the scale and prior mean parameters and an independent inverse-Wishart hyperprior distribution on the prior variance G centered at the identity matrix scaled by r with p degree of freedom. The same prior distribution is used for the ARS model, with the additional assumption that $\rho \sim U(0, 1)$ (non-negative autocorrelation between inclusion strata). In the LRS and NPRS models, $p(\sigma, \beta^*, \Lambda) \propto \sigma^{-2}$ and $p(\sigma, \beta^*, \tau) \propto \sigma^{-2}$ (standard non-informative scale prior distribution and hyperprior distribution). Description of the conditional draws of the Gibbs sampler are available at <http://www.sph.umich.edu/mrelliot/trim/meth2.pdf>.

The degree of compromise is a function of the mean and variance structure of the chosen model. The XRS and ARS models assume exchangeable slope means; the ARS model is more flexible in that its variance structure allows units with more nearly equal probabilities of inclusion to be smoothed more heavily than units with very unequal probabilities of inclusion. The LRS model assumes an underlying linear trend in slopes, whereas the NPRS model assumes only an underlying trend smooth up to its second derivative. Note that, in the LRS and NPRS models, we assume *a priori* independence for the regression parameters associated with a given covariate, i.e., $(\beta_{1j}, \dots, \beta_{Hj}) \perp (\beta_{1j'}, \dots, \beta_{Hj'})$, $j \neq j'$. This is because we model trends in these parameters across the inclusion stratum, and do not wish to "link up" these trends across the covariates.

Shrinkage will be greatest, corresponding to the most severe weight trimming, when the weight stratum slopes have little variability, or when the lowest probability-of-inclusion stratum are poorly estimated. Little shrinkage should occur when weight stratum slopes are precisely estimated and when they are systematically associated with their probability of inclusion. Based on Elliott and Little (2000), we would expect the XRS model to be the most efficient when large amounts of weight trimming are required to minimize MSE, but to be the most vulnerable to "overshrinking" when bias correction is most important. Increasing structure, particularly in the mean portion of the model as in LRS and NPRS, will provide more robust estimation in the sense that overshrinkage will occur only in near-pathological situations (e.g., when mean trends are

non-monotonic and highly discontinuous), and even then may only lead to slightly less bias correction than the data warrant. The price to be paid for this robustness, however, will be a reduction in efficiency relative to the exchangeable models.

3. Simulation results

Because we desire models that are simultaneously more efficient than design based estimators yet reasonably robust to model misspecification - and in general we feel that even Bayesian models should have good frequentist properties - we evaluate our proposed models in a repeated sampling context. We consider linear and logistic regression, under a misspecified model with a non-informative sampling design.

3.1 Linear regression

For the linear regression model in the presence of model misspecification, we generated population data as follows:

$$Y_i | X_i, \sigma^2 \sim N(\alpha X_i + \beta X_i^2, \sigma^2), \quad (11)$$

$$X_i \sim U(0, 10), i = 1, \dots, N = 20,000.$$

A noninformative, disproportionally stratified sampling scheme sampled elements as a function of X_i (I_i equals 1 if sampled and 0 otherwise):

$$h_i = \lceil X_i \rceil$$

$$P(I_i = 1 | h_i) = \pi_i \propto (1 + h_i/2.5) h_i$$

This created 10 strata, defined by the integer portions of the X_i values. Elements (Y_i, X_i) had $\approx 1/36^{\text{th}}$ the selection probability when $0 < X_i \leq 1$ as when $9 < X_i < 10$. We sampled $n = 500$ elements without replacement for each simulation. The object of the analysis is to obtain the population slope $B_1 = \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^N (X_i - \bar{X})^2$. We fixed $\alpha = \beta = 1$, yielding a positive bias in the estimate of B_1 , and varied σ^2 . The effect of model misspecification increases as $\sigma^2 \rightarrow 0$ as the bias of the estimators becomes larger relative to the variance, and conversely decreases as $\sigma^2 \rightarrow \infty$. We considered values of $\sigma^2 = 10^l$, $l = 1, \dots, 5$; 200 simulations were generated for each value of σ^2 .

Here and below we utilized an inverse-Wishart hyperprior distribution on the prior variance G , centered at the identity matrix with 2 degree of freedom.

In addition to the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) models

discussed in Section 2.3, we consider the standard designed-based (fully weighted) estimator, as well as trimmed weight and unweighted estimators. For the fully-weighted (FWT) estimator, we use the PMLE $B_w = (X'WX)^{-1} X'WY$ where, denoting by lower case the sampled elements ($I_i = 1$), $w_{hi} \equiv w_h$ for $h = 1, \dots, H$, $i = 1, \dots, n_h$, $W = \text{diag}(w_{hi})$, $x_{hi} = (1 x_i)$, X_h contains the stacked rows of x'_{hi} and X contain the stacked matrices X_h . We obtained inference about \hat{B}_w via the standard Taylor Series approximation (Binder 1983):

$$\text{Var}(\hat{B}_w) = \hat{S}_{XX}^{-1} \hat{\Sigma}(\hat{B}_w) \hat{S}_{XX}^{-1}$$

where \hat{S} is a design-consistent estimator of the population total $\sum_{i=1}^N x'_i x_i$ given by $X'WX$ and $\hat{\Sigma}(\hat{B}_w)$ is a design-consistent estimate of the variance of the total $\sum_{i=1}^N e_i x_i$ where $e_i = y_i - x'_i B$ is the difference between the value of y_i and its estimated value under the true population slope B : $\hat{\Sigma}(\hat{B}_w) = \sum_h n_h / (n_h - 1) \sum_{i=1}^{n_h} (\tilde{x}_{hi} - \bar{\tilde{x}}_h)(\tilde{x}_{hi} - \bar{\tilde{x}}_h)$, where $\tilde{x}_{hi} = w_{hi} e_{hi} x_{hi}$ for $e_{hi} = y_{hi} - x'_{hi} \hat{B}_w$. We also consider the trimmed (TWT) estimator obtained by replacing the weights w_{hi} with trimmed values w'_{hi} that set the maximum normalized value to 3: $w'_{hi} = N \tilde{w}'_{hi} / \sum_{h=1}^H n_h \tilde{w}'_{hi}$, where $\tilde{w}'_{hi} = \min(w_{hi}, 3N/n)$, and the unweighted (UNWT) estimator obtained by fixing $w_{hi} = N/n$ for all h, i .

Table 1 shows the relative bias, root mean square error (RMSE), and nominal 95% coverage for the three design-based and four model-based estimators of the population slope (second component of \hat{B}) under consideration, as a function of the variance σ^2 .

The fully-weighted estimator of the population slope is essentially design-unbiased under model misspecification; the unweighted and trimmed estimators are biased. The

biases of the exchangeable and autoregressive models increase as variance increases, as these models trade unbiasedness of the fully-weighted estimator for the reduced variance of the unweighted estimator. The linear and nonparametric model were approximately unbiased.

The unweighted and trimmed weight estimators perform poorly with respect to MSE for small values of σ^2 , where the bias due to model misspecification is critical, and well for larger values of σ^2 , where the instability of the fully-weighted estimator is more important than bias reduction. The exchangeable model-based estimator has good RMSE properties for small and large values of σ^2 , with MSE reductions of over 30%, but oversmooths for intermediate degrees of model specification. The autoregressive model performance equals that of the exchangeable model for small and large values of σ^2 , but is largely protected against the oversmoothing of the exchangeable models at intermediate levels. The linear and nonparametric models essentially dominated the fully weighted estimators with respect to MSE under all of the simulations considered, although MSE reductions were only on the order of 10%.

The unweighted and trimmed estimators have poor coverage except when model misspecification is nearly absent. The failure of the bias-variance tradeoff for the exchangeable estimator in the presence of model misspecification is evident in the poor coverage of the estimator for intermediate values of σ^2 ; this effect is ameliorated, but not completely removed, for the autoregressive estimator. The linear and non-parametric estimators have good coverage when model misspecification is less important but undercover to some degree when model misspecification is more important.

Table 1
Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population linear regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface

Estimator	Relative bias (%)					RMSE relative to FWT					True Coverage				
	Variance log ₁₀					Variance log ₁₀					Variance log ₁₀				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
UNWT	21.5	21.8	22.2	20.8	22.3	12.1	4.57	1.76	0.75	0.67	0	0	6	78	92
FWT	0.0	0.1	1.4	1.6	-0.2	1	1	1	1	1	94	95	96	94	96
TWT	8.3	8.4	9.6	8.8	7.8	4.74	1.88	1.02	0.71	0.75	0	13	78	95	96
XRS	0.2	2.2	11.4	15.1	18.3	1.00	1.17	1.18	0.73	0.68	87	86	64	91	96
ARS	0.1	1.4	9.6	14.5	17.4	1.00	1.03	1.11	0.74	0.69	87	89	78	90	96
LRS	-0.2	-0.4	1.1	1.6	-0.3	0.99	0.91	0.91	0.91	0.93	85	91	96	95	94
NPRS	-0.1	-0.3	0.9	1.5	-0.4	0.89	0.90	0.95	0.90	0.95	86	92	96	94	94

3.2 Logistic regression

For the logistic regression model, we generated population data as follows:

$$P(Y_i = 1 | X_i) \sim B(\text{expit}(3.25 - 0.75X_i + \gamma X_i^2)), \quad (12)$$

$$X_i \sim U(0, 10), i = 1, \dots, N = 20,000.$$

where $B(p)$ is a Bernoulli distribution with probability of "success" p , $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. The object of the analysis is to obtain the logistic population regression slope, defined as the value B_1 in the equation $\sum_{i=1}^N (y_i - \text{expit}(B_0 + B_1 x_i)) \left(\frac{1}{x_i} \right) = 0$. An unequal probability of selection sampling scheme was implemented as described in the linear regression simulations. We consider values of $\gamma = 0, 0.0158, 0.0273, 0.0368, 0.0454$, corresponding to curvature measures of $K = 0, 0.02, 0.04, 0.06, 0.08$ at the midpoint 5 of the support for X , where $K(X; \gamma) = |2\gamma/[1 + (2\gamma X - 0.75)^2]^{3/2}|$; 200 simulations were generated for each value of γ . As in the linear regression simulations, elements were sampled without replacement with probability proportional to $(1 + h_i/2.5)h_i$; a total of 1,000 elements were sampled for each simulation. We again considered the PMLE-based the fully weighted (FTW), unweighted (UNWT), and trimmed weight estimator (TWT), along with the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators. Inference about the PMLE estimators is obtained via Taylor Series approximations (Binder 1983), as discussed in the previous section.

Table 2 shows the relative bias, RMSE relative to the RMSE of the fully-weighted estimator, and true coverage of the nominal 95% CIs or PPIs associated with each of the

seven estimators of the population slope (B) for different values of curvature K , corresponding to increased degrees of misspecification.

The undersampling of small values of X meant that the maximum likelihood estimator of B in the model misspecification setting was unbiased for $K = 0$ and biased downward for $K = 0.02, 0.04, 0.06$, and 0.08 unless the sample design was accounted for. The trimmed estimator's bias was intermediate between the unweighted and fully weighted estimator. The exchangeable estimator's bias was between the trimmed weight estimator and fully weighted estimator; the autoregressive estimator's bias between that of the exchangeable and fully weighted estimator; while the linear and nonparametric estimators were essentially unbiased.

The unweighted estimator had substantially improved MSE (40% reduction) when the linear slope model was approximately correctly specified, but failed with moderate to large degree of misspecification. The trimmed weight, autoregressive, and nonparametric estimators all dominated the standard fully-weighted estimator, and the exchangeable and linear estimators nearly so, over the range of simulations considered. The crude trimming estimator yielded up to 30% reduction in MSE, the nonparametric, exchangeable and autoregressive estimators reductions of up to 20-25%, and the linear estimator reductions of only 10% or less.

The unweighted estimator had poor coverage except when the linear slope model was correctly specified, or nearly so. The model-based estimators had generally good coverage properties when the linear model was correctly specified, with slight reductions in coverage when curvature was substantial.

Table 2

Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population logistic regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface

Estimator	Relative bias (%)					RMSE relative to FWT					True Coverage				
	Curvature K					Curvature K					Curvature K				
Estimator	0	0.02	0.04	0.06	0.08	0	0.02	0.04	0.06	0.08	0	0.02	0.04	0.06	0.08
UNWT	1.0	-4.9	-11.9	-21.6	-34.6	0.57	0.73	0.88	1.19	1.61	96	89	66	32	17
FWT	1.1	2.2	1.3	-0.3	1.6	1	1	1	1	1	95	94	90	94	94
TWT	0.5	-1.0	-3.5	-7.2	-12.1	0.70	0.77	0.77	0.78	0.95	98	97	94	84	92
XRS	1.3	-0.8	-1.9	-5.6	-8.7	0.75	0.82	0.85	0.88	1.02	97	94	92	91	90
ARS	1.3	-0.5	-2.2	-4.8	-7.5	0.78	0.85	0.84	0.84	0.95	94	92	90	92	90
LRS	0.8	1.7	1.5	-0.4	1.1	0.89	0.97	0.94	0.91	1.02	95	91	88	92	89
NPRS	0.3	1.5	1.1	0.9	0.5	0.87	0.88	0.87	0.80	0.90	95	92	88	94	96

4. Application: Estimation of injuries to children in compact extended-cab pickup trucks

The Partners for Child Passenger Safety dataset consists of the disproportionate, known-probability sample from all State Farm claims since December 1998 involving at least one child occupant ≤ 15 years of age riding in a model year 1990 or newer State Farm-insured vehicle (Durbin, Bhatia, Holmes, Shaw, Werner, Sorenson and Winston 2001). Because injuries, and especially “consequential” injuries defined as facial lacerations or other injuries rated 2 or more on the Abbreviated Injury Scale (AIS) (Association for the Advancement of Automotive Medicine 1990), are relatively rare even among children in the population of crash-related vehicle damage claims, a disproportional stratified cluster sample is used to select vehicles (the unit of sampling) for the conduct of a telephone survey with the driver. Vehicles containing children who received medical treatment following the crash were over-sampled so that the majority of injured children would be selected while maintaining the representativeness of the overall population. (Medical treatment is defined as treatment by paramedics, treatment at a physician’s office or emergency room, or hospitalization.) If a vehicle was sampled, all child occupants in that vehicle were included in the survey. Drivers of sampled vehicles were contacted by phone and, if medical treatment had been received by a passenger, screened via an abbreviated survey to verify the presence of at least one child occupant with an injury. All vehicles with at least one child who screened positive for injury and a 10% random sample of vehicles in which all child occupants who were reported to receive medical treatment but screened negative for injury were selected for a full interview; a 2% (later 2.5%) sample of crashes where no medical treatment was received were also selected. Because the treatment stratification is imperfectly associated with risk of injury (more than 15% of the population with consequential injuries are estimated to be in the lowest probability-of-selection category and nearly 20% of those without consequential injuries are in the highest probability-of-selection category), the sampling design is informative, with unweighted odds ratios biased toward the null (Korn and Graubard 1995). In addition, the weights for this dataset are quite variable: $1 \leq w_i \leq 50$, where 9% of the weights have normalized values greater than 3.

Winston, Kallan, Elliott, Menon and Durbin (2002) determined that children rear-seated in compacted extended cab pickups are at greater risk of consequential injuries than children rear-seated in other vehicles. However, quantifying degree of excess risk, and thus the size of the public health problem, was problematic. The unweighted odds ratio (OR)

of consequential injury for children riding in compacted extended cab pickups versus other vehicles was 3.54 (95% CI 2.01, 6.23), versus the fully-weighted estimator of 11.32 (95% CI 2.67, 48.03). Because both injury risk and compacted extended cab pickup use were associated with child age, crash severity (passenger compartment intrusion and drivability), direction of impact, and vehicle weight, a multivariate logistic regression model that adjusted for these factors was also considered. The unweighted and fully-weighted adjusted ORs for injury risk in rear seated children in compacted extended cab pickups versus other vehicles are 3.50 (95% CI 1.88, 6.53) and 14.56 (95% CI 3.45, 61.40) respectively. Utilizing the unweighted estimator was problematic because of bias toward the null induced by the informative sample design; however the fully weighted estimator appeared to be highly unstable, in part because of the presence of one consequential-injured child in the compact extended cab pickups had a very low probability of selection (0.025). In Winston *et al.* (2002), this child was removed before conducting the analysis.

Table 3 shows the results for the unadjusted and adjusted odds ratios of consequential injury risk using the unweighted, fully-weighted, and trimmed-weight design-based estimators, along with the model-based exchangeable, autoregressive, and linear regression slope models. (Results for the model-based estimators from 250,000 draws of a single chain after a 50,000 draw burn-in; convergence was assessed via Geweke (1992).) For the XRS and ARS models, $p(\Sigma) \sim \text{INVERSE-WISHART}(p, 0.1I)$, where $p=2$ for the unadjusted model and $p=13$ for the adjusted model. In the unadjusted results, the XRS and ARS estimators are intermediate between the unweighted and fully-weighted estimator, while the linear and nonparametric estimators tends to track the fully-weighted estimator. In the adjusted analysis, all three model-based estimators are intermediate between the unweighted and fully-weighted estimators, with the XRS estimator closest to the unweighted estimator and the LRS estimator closest to the fully-weighted estimator. Based on the results of the simulation, it appears that the ARS estimator, which suggest relative risks of injury on the order of 7 for children in compact extended cab pickups relative to other vehicles, may be a better estimator of relative risk than either the unweighted or fully weighted estimator. (As a “sanity check” of sorts, we note that an additional two years of data, not available at the time of Winston *et al.* (2002), which included an additional 4,091 rear-seated children in passenger vehicles [44 in compact extended-cab pickup trucks], provided a fully-weighted unadjusted odds ratio for injury for children in compact-extended cab pickups of 6.3, and an adjusted OR of 7.0.)

Table 3

Estimated odds ratio of injury for children rear-seated in compacted extended cab pickups ($n = 60$) versus rear-seated in other vehicles ($n = 8,060$), using unweighted (UNWT), fully-weighted (FWT), weights trimmed to a normalized value of 3 (TWT), exchangeable random slope (XRS), autoregression random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators; unadjusted and adjusted for child age, crash severity, direction of impact, and vehicle weight. Point estimates for XRS, ARS, and LRS models from posterior median. 95% confidence interval or posterior predictive interval in subscript. Data from Partners for Child Passenger Safety

	UNWT	FWT	TWT	
Unadj.	3.54 (2.01, 6.23)	11.32 (2.67, 48.02)	9.15 (2.65, 31.57)	
Adj.	3.50 (1.88, 6.53)	14.56 (3.45, 61.40)	10.99 (2.97, 34.64)	
	XRS	ARS	LRS	NPRS
Unadj.	6.70 (2.51, 20.92)	6.69 (2.64, 21.05)	11.17 (3.21, 24.94)	10.34 (3.27, 24.62)
Adj.	4.45 (2.39, 8.67)	6.67 (3.56, 11.94)	11.87 (3.33, 36.93)	10.23 (3.02, 37.93)

5. Discussion

The models discussed in this paper generalize the work of Lazzeroni and Little (1998) and Elliott and Little (2000), where population inference was restricted to population means under Gaussian distributional assumptions. Viewing weighting as an interaction between inclusion probability and model parameters opens up an alternative paradigm for weight trimming as a random effects model that smoothes model parameters of interest across inclusion classes. Models with exchangeable mean structures offer the largest degree of shrinkage or trimming but the most sensitivity to model misspecification; models with highly structured means are potentially less efficient but are more robust to model misspecification. This robustness property may be particularly important in light of the fact that elements of the large inclusion strata provide the largest degree of potential variance reduction in the model-based setting but are also subject to the largest degree of model bias and variance due to extrapolation.

We consider simulations under varying degrees of model misspecification and informative sampling for both linear and logistic regression models. The linear and non-parametric smoothing models nearly dominated fully-weighted estimators with respect to squared error loss in the simulations considered. The exchangeable model showed some tendency to oversmooth, favoring variance reduction over bias correction, especially in the linear regression setting. All of the weight smoothing estimators tended to have less than nominal coverage when models were highly misspecified, although in no case was the nominal coverage catastrophically low. The autoregressive smoothing model, which allows for differential degrees of local smoothing across weight strata, appeared to provide non-trivial

increases in efficiency with limited risk of severe over-smoothing or undercoverage.

Applying the methods to the Partners for Child Passenger Safety data to determine the excess risk of injury in a crash to rear-seated children in compacted extended-cab pickups relative to rear-seated children in other passenger vehicles, it appears that the decision in Winston *et al.* (2002) to eliminate a low probability-of-selection child from the analysis to stabilize the estimates was indeed conservative. Indeed, the ARS estimator, favored by MSE in simulations, suggests an adjusted excess risk of 6.7 with a 95% PPI of (3.6, 11.9), versus the 14.6 with 95% CI of (3.4, 61.4) of the fully-weighted estimator.

Although this paper utilizes a fully Bayesian approach to inference about the posterior predictive distribution of the population regression slope, empirical Bayes (EB) estimates can also be obtained via ML or REML estimation using standard linear or generalized linear mixed model methods. In the Gaussian setting, the EB estimates of G and σ^2 can be “plugged into” the closed-form expressions for $E(\mathbf{B}|y, X)$ and $\text{Var}(\mathbf{B}|y, X)$. The general exponential setting is more problematic. The plug-in estimates can be used to determine $E(\mathbf{B}|y, X)$ via root-finding methods; the lack of a closed form for $E(\mathbf{B}|y, X)$ makes it difficult to obtain model-based Empirical Bayes estimators for $\text{Var}(\mathbf{B}|y, X)$. Also, standard Empirical Bayes estimators do not account for the uncertainty in the estimation of G .

We also note that, while computation of the actual trimming values of the case weights is unnecessary in this approach, it is possible to determine the revised design weights implied by the shrinkage. In the linear model setting, these can be obtained via a iterative application of a calibration weighting scheme such as generalized regression estimators or GREG (Deville and Särndal 1992). The

general exponential setting required embedding the calibration weight algorithm within the iterative reweighted least squares (IRWLS) algorithm used to fit a generalized linear model.

When sampling weights are used to account for misspecification of the mean in a regression setting, it could be argued that the correct approach is to correctly specify the mean to eliminate discrepancies between the fully-weighted and unweighted estimates of the regression parameters. However, perfect specification is an unattainable goal, and even good approximations might be highly biased if case weights are ignored when the sampling probabilities are highly variable. In the informative sampling setting, it may be impossible to determine whether discrepancies between weighted and unweighted estimates are due to model misspecification or to the sample design itself. Finally, even misspecified regression models have the attractive feature in the finite population setting of yielding a unique target population quantity. Consequently accounting for the probability of inclusion in linear and generalized linear model settings continues to be advised, and methods that balance between a low-bias, high variance fully-weighted analysis and a high bias, low variance unweighted analysis remain useful.

The methods discussed in this paper show the promise of adapting model-based methods to attack problems in survey data analysis. Our goal is not to develop a single hierarchical Bayesian model finely-tuned to a specific or question dataset at hand, but to develop robust yet efficient methods that can be applied in a fast-paced "automated" setting that many applied survey research analysts must sometimes work. Although computationally intensive, the methods considered are applications or extensions of the existing random-effect model "toolbox," and can either be implemented in existing statistical packages or executed with relatively simple MCMC methods. Our approach retains a design-based flavor in that we attempt to develop "automated" Bayesian model-based estimation techniques that yield robust inference in a repeated-sampling setting when the model itself is misspecified. However, because these models rely on stratifying the data by probability of selection as a prelude to using pooling or shrinkage techniques to induce data-driven weight trimming, there is a natural correspondence between this methodology and (post)stratified sample designs in which strata correspond to unequal probabilities of inclusion. Developing methods that accommodate a more general class of complex sample designs that include single or multi-stage cluster samples and/or strata that "cross" the inclusion strata remains an important area for future work.

Acknowledgements

The author thanks Roderick J.A. Little, along with the Editor, Associate Editor, and two anonymous reviewers, for their review and comments. The author also thanks Drs. Dennis Durbin and Flaura Winston of the Partners for Child Passenger Safety project for their assistance, as well as State Farm Insurance Companies for their support of the Partners for Child Passenger Safety project. This research was supported by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01.

References

- Alexander, C.H., Dahl, S. and Weidman, L. (1997). Making estimates from the American Community Survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 2000, 88-97.
- Association for the Advancement of Automotive Medicine (1990). *The Abbreviated Injury Scale, 1990 Revision*. Association for the Advancement of Automotive Medicine, Des Plaines, Illinois.
- Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195-208.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Durbin, D.R., Bhatia, E., Holmes, J.H., Shaw, K.N., Werner, J.V., Sorenson, W. and Winston, F.K. (2001). Partners for child passenger safety: A unique child-specific crash surveillance system. *Accident Analysis and Prevention*, 33, 407-412.
- Elliott, M.R., and Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.
- Ericson, W.A. (1969). Subjective bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-234.
- Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2000, 598-603.
- Gelfand, A.E., and Smith, A.M.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389-409.
- Gelman, A., and Carlin, J.B. (2002). Poststratification and weighting adjustments. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), 289-302.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), 89-193.

- Ghosh, M., and Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.
- Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*, London: Chapman and Hall.
- Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Kish, L. (1992). Weighting for unequal P_i . *Journal of Official Statistics*, 8, 183-200.
- Korn, E.L., and Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.
- Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.
- Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society, Series B*, 65, 175-190.
- Lazzeroni, L.C., and Little, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.
- Lin, X., and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61, 381-400.
- Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.
- Little, R.J.A. (1993). Poststratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A. (2004). To model or not model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition. CRC Press: Boca Raton, Florida.
- Oh, H.L., and Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse, *Incomplete Data in Sample Surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184.
- Pfaffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Pfaffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rabash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.
- Potter, F. (1990). A study of procedures to identify and trim extreme sample weights. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1990, 225-230.
- Rizzo, L. (1992). Conditionally consistent estimators using only probabilities of selection in complex sample surveys. *Journal of the American Statistical Association*, 87, 1166-1173.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: John Wiley & Sons, Inc.
- Särndal, C.-E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.
- Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, 40, 364-372.
- Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93, 341-348.
- Winston, F.K., Kallan, M.K., Elliott, M.R., Menon, R.A. and Durbin, D.R. (2002). Risk of injury to child passengers in compact extended pick-up trucks. *Journal of the American Medical Association*, 287, 1147-1152.

Semiparametric model-assisted estimation for natural resource surveys

F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson and M. Giovanna Ranalli¹

Abstract

Auxiliary information is often used to improve the precision of survey estimators of finite population means and totals through ratio or linear regression estimation techniques. Resulting estimators have good theoretical and practical properties, including invariance, calibration and design consistency. However, it is not always clear that ratio or linear models are good approximations to the true relationship between the auxiliary variables and the variable of interest in the survey, resulting in efficiency loss when the model is not appropriate. In this article, we explain how regression estimation can be extended to incorporate semiparametric regression models, in both simple and more complicated designs. While maintaining the good theoretical and practical properties of the linear models, semiparametric models are better able to capture complicated relationships between variables. This often results in substantial gains in efficiency. The applicability of the approach for complex designs using multiple types of auxiliary variables will be illustrated by estimating several acidification-related characteristics for a survey of lakes in the Northeastern US.

Key Words: Regression estimation; Smoothing; Kernel regression; Lake chemistry.

1. Introduction

Post-stratification, calibration and regression estimation are different design-based approaches that can be used to improve the precision of estimators when auxiliary information is available at the estimation stage. *Model-assisted estimation* (Särndal, Swensson and Wretman 1992) provides a convenient framework in which to develop these and related survey estimators. Under that framework, a superpopulation model describes the relationship between the variable of interest and the auxiliary variables. This model is then used to construct sample-based estimators that have improved precision when the model is correct, but maintain key design properties such as consistency and an estimable variance when the model is incorrect.

Until recently, the superpopulation models used in this context were formulated as parametric models, most often ratio or linear models. While reasonable in many practical applications, there are also many situations in which such relatively simple models are not good representations of the relationship between the variable of interest and the auxiliary variables. In Breidt and Opsomer (2000), a nonparametric model-assisted estimator was proposed based on local polynomial regression, which generalized the well-established parametric regression estimators. With this estimator, the superpopulation is no longer required to follow a pre-specified parametric shape. Instead, the relationship between the the variable(s) of interest in the survey and the auxiliary variable is required to be smooth (continuous), but is otherwise left completely unspecified.

In the current paper, we formally extend the theory of Breidt and Opsomer (2000) to the semiparametric regression context, in which some variables are incorporated linearly, and others are incorporated through smooth additive terms. This extension makes their results more useful in practice, since auxiliary information is very often multi-dimensional in nature, and almost always contains categorical variables that need to enter the regression model parametrically (through the use of indicator variables). An illustration of this is provided by a survey of lakes in the Northeastern states of the U.S. conducted by the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency. In that survey, 334 lakes were sampled from a population of 21,026 lakes between 1991 and 1996. We will apply the semiparametric model-assisted estimator to produce estimates of the mean and distribution function of the *acid neutralizing capacity* and other chemistry variables of interest. In this application, we will include in the model both categorical and continuous variables linearly and a continuous variable as a smooth additive term.

In Opsomer, Breidt, Moisen and Kauermann (2007), the nonparametric model-assisted estimation principle was extended to generalized additive models (GAMs) and applied in an interaction model for the estimation of variables from Forest Inventory and Analysis surveys. While GAMs also contained a mixture of categorical (parametric) and nonparametric terms, a complete theoretical development is not possible in the case of GAMs, and was therefore not provided there. The semiparametric model considered in this article can be viewed as a special case of a GAM with an identity link function. Unlike the

1. F. Jay Breidt, Department of Statistics, Colorado State University, Fort Collins CO 80523, U.S.A.; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. E-Mail: jopsomer@iastate.edu; Alicia A. Johnson, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis MN 55455, U.S.A.; M. Giovanna Ranalli, Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli, 06123 Perugia, Italy.

“general” GAM, the semiparametric model allows for formal derivation of the statistical properties of the model-assisted estimator.

The remainder of the article is structured as follows. In Section 2, the semiparametric model-assisted estimator is defined. Section 3 states and proves the design properties of the estimator. Section 4 describes the application of semiparametric model-assisted estimation to the Northeastern Lakes data. Section 5 provides a conclusion.

2. Semiparametric model-assisted estimation

We begin by considering the superpopulation model with a single univariate nonparametric term and a parametric component; extension to several nonparametric terms is addressed in Section 3.2. The parametric component can be composed of an arbitrary number of linear terms. This model is the semiparametric model studied by Speckman (1988), among others. This superpopulation model, which we denote by ξ , can be written down as

$$\begin{aligned} E_\xi(y_k) &= g(x_k, \mathbf{z}_k) = m(x_k) + \mathbf{z}_k \boldsymbol{\beta} \\ \text{Var}_\xi(y_k) &= v(x_k, \mathbf{z}_k) \end{aligned} \quad (1)$$

with x_k a continuous auxiliary variable to be modelled nonparametrically and $\mathbf{z}_k = (z_{1k}, \dots, z_{Dk})$ a vector of D categorical or continuous auxiliary variables that are parametrically specified. The functions $m(\cdot)$ and $v(\cdot, \cdot)$ and the parameter vector $\boldsymbol{\beta}$ are unknown. For identifiability purposes, we will assume that the vector \mathbf{z}_k contains an intercept term, and that the function $m(\cdot)$ is centered around 0 with respect to the distribution of the x_k . We will derive the model-assisted estimator that uses model (1) by first defining population-level estimators for the unknown functions and parameters, and then constructing sample-based estimators. This is the same approach used for the parametric case in Särndal *et al.* (1992, Chapter 6).

Let $U = \{1, 2, \dots, N\}$ represent the ordered labels for a finite population of interest. As the population estimator for $g(x_k, \mathbf{z}_k)$, we will use the *backfitting estimator* described in Opsomer and Ruppert (1999). We first introduce the required notation. Let $K(\cdot)$ represent a kernel function used to define the neighborhoods in which the local polynomials will be fitted (assumptions on K are specified in the Appendix). The population *smoother vector* for local polynomial regression of degree p at x_k is defined as

$$\mathbf{s}_{Uk}^T = \mathbf{e}_1^T (\mathbf{X}_{Uk}^T \mathbf{W}_{Uk} \mathbf{X}_{Uk})^{-1} \mathbf{X}_{Uk}^T \mathbf{W}_{Uk}$$

with \mathbf{e}_1 a vector of length $p+1$ with a 1 in the first position and 0s elsewhere, $\mathbf{W}_{Uk} = \text{diag}\{h^{-1}K((x_1 - x_k)/h), \dots, h^{-1}K((x_N - x_k)/h)\}$ and

$$\mathbf{X}_{Uk} = \begin{bmatrix} 1 & x_1 - x_k & \dots & (x_1 - x_k)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N - x_k & \dots & (x_N - x_k)^p \end{bmatrix}.$$

The smoother \mathbf{s}_{Uk} can be applied to the vector $\mathbf{Y}_U = (y_1, \dots, y_N)^T$ to produce the nonparametric regression fit with respect to the variable x at observation x_k . It can also be applied to any of the columns of $\mathbf{Z}_U = (\mathbf{z}_1^T, \dots, \mathbf{z}_N^T)^T$ to smooth those with respect to x . This will be done in the derivation of the properties of the semiparametric estimator (Section 3).

In addition to the smoother vector at x_k , \mathbf{s}_{Uk}^T , we also need to define the *smoother matrix* at all the observation points x_1, \dots, x_N ,

$$\mathbf{S}_U = \begin{bmatrix} \mathbf{s}_{U1}^T \\ \vdots \\ \mathbf{s}_{UN}^T \end{bmatrix},$$

and the *centered smoother matrix* $\mathbf{S}_U^* = (\mathbf{I} - \mathbf{1}\mathbf{1}^T/N)\mathbf{S}_U$. When the smoother matrix is applied to \mathbf{Y}_U , it produces the vector of nonparametric regression fits at all the observation points. The centered smoother matrix \mathbf{S}_U^* produces centered fits, *i.e.*, the overall mean of the fitted values is subtracted from each fitted value. The centering is used to maintain identifiability of the estimators, as explained in Opsomer and Ruppert (1999).

For any observation x_k , a possible estimator of $m(x_k)$ could be defined as $\mathbf{s}_{Uk}^T \mathbf{Y}_U$, with or without a centering adjustment. This estimator would generally be poor, since it does not take into account the fact that the y_k contain a parametric component that depends on the \mathbf{z}_k . A more efficient estimator is provided by jointly estimating both $m(\cdot)$ and $\boldsymbol{\beta}$, as is done by the following set of estimators

$$\begin{aligned} \mathbf{B} &= (\mathbf{Z}_U^T (\mathbf{I} - \mathbf{S}_U^*) \mathbf{Z}_U)^{-1} \mathbf{Z}_U^T (\mathbf{I} - \mathbf{S}_U^*) \mathbf{Y}_U \\ m_k &= \mathbf{s}_{Uk}^T (\mathbf{Y}_U - \mathbf{Z}_U \mathbf{B}) \quad k=1, \dots, N. \end{aligned} \quad (2)$$

In these estimators, \mathbf{B} is calculated first, and then the “residual vector” $\mathbf{Y}_U - \mathbf{Z}_U \mathbf{B}$ is smoothed with respect to x . The estimators in (2) are identical to the *backfitting estimators* for additive models described in Hastie and Tibshirani (1990) and implemented in *gam* in S-Plus, R or SAS. As a population estimator for $E_\xi(y_k) = g(x|k, \mathbf{z}_k)$, we use

$$g_k = m_k + \mathbf{z}_k \boldsymbol{\beta}.$$

We now explain how to construct a model-assisted estimator based on the semiparametric regression approach. Let $A \subset U$ be a sample of size n drawn from U according to sampling design $p(A)$ with one-way and two-way

inclusion probabilities $\pi_k = \sum_{A \ni k} p(A)$, $\pi_{kl} = \sum_{A \ni k, l} p(A)$, respectively. If the g_k , $k = 1, \dots, N$ were available, it would be possible to construct a *difference estimator* for the population mean of the y_k , $\bar{y}_N = \sum_U y_k / N$, as

$$\hat{y}_{\text{dif}} = \frac{1}{N} \sum_U g_k + \frac{1}{N} \sum_A \frac{y_k - g_k}{\pi_k}, \quad (3)$$

which is design unbiased and has design variance

$$\text{Var}_p(\hat{y}_{\text{dif}}) = \frac{1}{N^2} \sum_U \sum_{\pi_{kl}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k - g_k}{\pi_k} \frac{y_l - g_l}{\pi_l}$$

(Särndal *et al.* 1992, page 221). The design variance is small if the deviations between y_k and g_k are small. This estimator is not feasible, since it requires knowledge of all the x_k , z_k and y_k for the population to calculate. Instead, we will construct a feasible estimator by replacing the g_k by sample-based estimators. The sample-based estimators corresponding to the population estimators in (2) are constructed as follows. The design-weighted local polynomial smoother vector is

$$s_{Ak}^{0T} = e_1^T (X_{Ak}^T W_{Ak} X_{Ak})^{-1} X_{Ak}^T W_{Ak}, \quad (4)$$

with X_{Ak} containing the rows of X_{Uk} corresponding to the $k \in A$ and

$$W_{Ak} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left(\frac{x_j - x_k}{h} \right) : j \in A \right\}.$$

The matrix $X_{Ak}^T W_{Ak} X_{Ak}$ in (4) will be singular if, for some sample A , there are less than $p+1$ observations in the support of the kernel at some x_k . This issue can be avoided in practice by selecting a bandwidth large enough to make that matrix invertible. However, this situation cannot be excluded in general and we need an estimator that exists for every sample A for the theoretical derivations of Section 3. Hence, we will consider the following adjusted sample smoother vector

$$s_{Ak}^T = e_1^T (X_{Ak}^T W_{Ak} X_{Ak} + \text{diag}(\delta N^{-2}))^{-1} X_{Ak}^T W_{Ak}, \quad (5)$$

for some small $\delta > 0$, as done in Breidt and Opsomer (2000). The sample smoother matrix and its centered version are

$$S_A = [s_{Ak}^T : k \in A] \quad S_A^* = (I - 11^T \Pi_A^{-1} / N) S_A$$

with $\Pi_A = \text{diag}\{\pi_k : k \in A\}$. The design-weighted estimators for B and the m_k are

$$\hat{B} = (Z_A^T \Pi_A^{-1} (I - S_A^*) Z_A)^{-1} Z_A^T \Pi_A^{-1} (I - S_A^*) Y_A \quad (6)$$

$$\hat{m}_k = s_{Ak}^T (Y_A - Z_A^T \hat{B}), \quad (7)$$

where Z_A and Y_A denote the sample versions of Z_U and Y_U , respectively. Note that the estimator \hat{m}_k is defined for any x_k in the population, not only those appearing in the sample. As for the population estimators, these estimators can again be written as the solution to backfitting equations, so that they can be calculated by appropriately weighted versions of the existing algorithms. The estimator for g_k is

$$\hat{g}_k = \hat{m}_k + z_k \hat{B}.$$

The semiparametric model-assisted estimator is then constructed by replacing the g_k in (3) by the \hat{g}_k :

$$\hat{y}_{\text{reg}} = \frac{1}{N} \sum_U \hat{g}_k + \frac{1}{N} \sum_A \frac{y_k - \hat{g}_k}{\pi_k}. \quad (8)$$

Defining $\bar{y}_\pi = \sum_A y_k / \pi_k$ and similarly for \bar{z}_π , an equivalent expression for \hat{y}_{reg} is given by

$$\hat{y}_{\text{reg}} = \bar{y}_\pi + (\bar{z}_\pi - \bar{z}_N) \hat{B} + \frac{1}{N} \sum_U \hat{m}_k - \frac{1}{N} \sum_A \frac{\hat{m}_k}{\pi_k}, \quad (9)$$

which shows that the semiparametric estimator can be interpreted as a “traditional” linear regression survey estimator using the parametric model component $z\beta$, with an additional correction term for the nonparametric component of the model. This estimator also shares some desirable properties with the fully parametric regression estimators. It is location and scale invariant, and it is calibrated for both the parametric and the nonparametric model components, in the sense that $\hat{x}_{\text{reg}} = \bar{x}_N$ and $\hat{z}_{\text{reg}} = \bar{z}_N$. The calibration for the variables in the parametric term can be checked directly by using expressions (6) and (7), while the calibration for the nonparametrically specified variable x_k follows from the fact that $s_{Ak}^T X_A = x_k$, where $X_A = (x_k : k \in A)^T$ (we are ignoring the effect of the adjustment $\text{diag}(\delta N^{-2})$ in (5), because that adjustment can be made arbitrarily small). In addition, the estimator can be written as a weighted sum of the y_k , $k \in A$, so that a set of weights w_k can be obtained and applied to any survey variable of interest.

3. Properties and extensions

3.1 Design properties

In this section, we explore the design properties of the semiparametric estimator (8). In particular, we prove that \hat{y}_{reg} is design \sqrt{n} -consistent, and we derive its asymptotic distribution, including an estimated variance. This will be done in the design-asymptotic context used in Isaki and Fuller (1982) and in Breidt and Opsomer (2000), in which both the population and the samples increase in size as $N \rightarrow \infty$. All proofs and the necessary assumptions are in the Appendix.

In the following theorem, we prove the design consistency of the semiparametric estimator. We also show that the convergence rate is \sqrt{n} , the usual rate for design estimators.

Theorem 3.1 *Under the assumptions A1–A8, the estimator \hat{y}_{reg} in (8) is design consistent with rate \sqrt{n} , in the sense that*

$$\hat{y}_{\text{reg}} = \bar{y}_N + O_p\left(\frac{1}{\sqrt{n}}\right).$$

The following theorem proves that a central limit theorem for \hat{y}_{reg} exists whenever it exists for the expansion estimator \bar{y}_π .

Theorem 3.2 *Under the assumptions A1–A8, if*

$$\frac{\bar{y}_\pi - \bar{y}_N}{\sqrt{\hat{V}(\bar{y}_\pi)}} \rightarrow N(0, 1),$$

with

$$\hat{V}(\bar{y}_\pi) = \frac{1}{N^2} \sum_A \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k y_l}{\pi_k \pi_l}$$

for a given sampling design, then we also have

$$\frac{\hat{y}_{\text{reg}} - \bar{y}_N}{\sqrt{\hat{V}(\hat{y}_{\text{reg}})}} \rightarrow N(0, 1),$$

with

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum_A \sum_A \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - \hat{g}_k}{\pi_k} \frac{y_l - \hat{g}_l}{\pi_l}. \quad (10)$$

3.2 Semiparametric additive model

The results in Theorems 3.1 and 3.2 use the semiparametric model (1), which contains a single univariate nonparametric term $m(\cdot)$. In many practical applications, several auxiliary variables will be available that could be included in the nonparametric portion of a model, but the curse of dimensionality makes it often difficult to combine several variables into a single multi-dimensional nonparametric term. Instead, the variables that are to be included nonparametrically will be treated as univariate components. This results in the *semiparametric additive model*, which is written as

$$E_\xi(y_k) = g(x_k, z_k) = m_1(x_{1k}) + \dots + m_Q(x_{Qk}) + z_k \beta$$

$$\text{Var}_\xi(y_k) = v(x_k, z_k)$$

where the $m_q(\cdot)$, $q=1, \dots, Q$ and $v(\cdot, \cdot)$ are unknown smooth functions.

When $Q=2$, expressions similar to (6) and (7) can be developed, using the additive model decompositions of

Opsomer and Ruppert (1997), and for $Q>2$, recursive expressions can be derived using the approach of Opsomer (2000). The estimator would then be written as in equations (6) and (7), but with the smoother vectors s_{Ak} and smoother matrix S_A replaced by complicated higher-dimensional additive model smoothers (see Opsomer (2000) for details). Because of this, formally proving the properties of the model-assisted estimator for the case with arbitrary Q would be a challenging task beyond the scope of the current article.

In practice, the backfitting algorithm formulation provides a much more efficient and simple way to calculate the semiparametric estimator. Let s_{Aqk} represent the sample smoother vector, as defined in (5), for the variable x_q at the observation x_{qk} and S_{Aq} is the corresponding smoother matrix for the variable x_q . Also, \hat{m}_{qk} denotes the sample-weighted backfitting estimator for $m_q(x_{qk})$ and $\hat{m}_{Aq} = (\hat{m}_{qk}, k \in A)$. The backfitting algorithm for a model including Q nonparametric terms consists of the following set of equations, iterated to converge:

$$\hat{B} = (Z_A^T \Pi_A^{-1} Z_A)^{-1} Z_A^T \Pi_A^{-1} \left(Y_A - \sum_{q=1}^Q \hat{m}_{Aq} \right)$$

$$\hat{m}_{A1} = S_{A1} \left(Y_A - Z_A^T \hat{B} - \sum_{q \neq 1} \hat{m}_{Aq} \right)$$

$$\vdots$$

$$\hat{m}_{AQ} = S_{AQ} \left(Y_A - Z_A^T \hat{B} - \sum_{q \neq Q} \hat{m}_{Aq} \right).$$

These equations provide weighted fits at the sample locations $k \in A$ only. For the remaining locations $k \in U$ not in A , an additional smoothing step is required after obtaining the \hat{m}_{Aq} , $q=1, \dots, Q$:

$$\hat{m}_{kq} = s_{Aqk}^T \left(Y_A - Z_A^T \hat{B} - \sum_{q \neq q} \hat{m}_{Aq} \right).$$

The sample-based estimators for the mean function at all $k \in U$ are then defined as $\hat{g}_k = \hat{m}_{k1} + \dots + \hat{m}_{kQ} + z_k \hat{B}$, which are used in expression (8) to construct the model-assisted estimator.

4. Application to Northeastern Lakes survey

In this section, we will show the applicability of the semiparametric regression estimator on a dataset of water chemistry samples. As will be illustrated, once a set of auxiliary variables and a model has been selected, computing survey estimators for the semiparametric model is as easy as for linear models, and hence can lead to improved precision for relatively little cost.

The National Surface Water Survey (NSWS) sponsored by the U.S. Environmental Protection Agency (EPA) between the years of 1984 and 1986 estimated 4.2 percent of the lakes in the northeastern region of the United States to be acidic (Stoddard, Kahl, Deviney, DeWalle, Driscoll, Herlihy, Kellogg, Murdoch, Webb and Webster 2003). Acid-sensitive Northeastern lakes were among the concerns addressed by the Clean Air Act Amendment (CAAA) of 1990, which placed restrictions on industrial sulfur and nitrogen emissions in an effort to reduce the acidity of these waters. A common measurement of acidity is acid neutralizing capacity (ANC), which is defined as a water's ability to buffer acid. An ANC value less than zero $\mu\text{eq}/L$ indicates that the water has lost all ability to buffer acid. Surface waters with ANC values below 200 $\mu\text{eq}/L$ are considered at risk of acidification, and values less than 50 $\mu\text{eq}/L$ are considered at high risk (National Acid Precipitation Assessment Program (1991), page 15).

Between 1991 and 1996, the Environmental Monitoring and Assessment Program (EMAP) of the U.S. Environmental Protection Agency conducted a survey of lakes in the Northeastern states of the U.S. These data were collected in order to determine the effect that restrictions put in place by the CAAA had on the ecological condition of these waters. The survey is based on a population of 21,026 lakes from which 334 lakes were surveyed, some of which were visited several times during the study period. Multiple measurements on the same lake were averaged in order to obtain one measurement per lake sampled. Lakes to be included in the survey were selected using a complex sampling design commonly employed by EMAP based on a hexagonal grid frame (see Larsen, Thornton, Urquhart and Paulsen (1993) for a description of the sampling design).

Let y_k represent the (possibly averaged) ANC value of the k^{th} sampled lake. A very simple estimate of the ANC mean of the lakes is represented by the expansion estimator \bar{y}_π . In this as in many surveys, a better choice is the Hájek estimator,

$$\hat{y}_H = \frac{1}{\hat{N}} \sum_{k \in A} \frac{y_k}{\pi_k}, \quad (11)$$

which applies a ratio type adjustment for the estimation of the population size through $\hat{N} = \sum_{k \in A} 1/\pi_k$. However, auxiliary variables are available for each lake in this population, so that it should be possible to further improve upon the efficiency of the Hájek estimator. The following variables are available for each $k \in U$:

x_k = UTMX, x -geographical coordinate of the centroid of each lake in the UTM coordinate system,

$z_{j,k}$ = indicator variable for eco-region $j = 1, \dots, 6$,

$z_{7,k}$ = UTMY, y -geographical coordinate,

$z_{8,k}$ = elevation.

There are seven different eco-regions included in the population, thus dummy variables $z_{j,k}$ are constructed for $j = 1, \dots, 6$. A semiparametric regression estimator for the variable y will be constructed by treating the UTMX variable x as a nonparametric term and the remaining variables $z_1 - z_8$ as a parametric component. Model selection was used to determine that treating the other two continuous variables as nonparametric did not improve the model fit. For comparison purposes, we also computed a regression estimator that treats all terms as parametric. This estimator is therefore identical to the semiparametric estimator, except that the x -geographical coordinate is modeled linearly. We will denote this fully parametric regression estimator by \hat{y}_{par} .

In order to determine the estimated efficiency of survey estimators, we need to compute the variance estimates. However, second order inclusion probabilities were not available, thus we cannot evaluate $\hat{V}(\hat{y}_{\text{reg}})$ as in (10). In order to come up with appropriate variance estimates, we treat the complex sampling design as a stratified sample taken with replacement. The 14 strata we selected correspond to groups of spatial clusters of lakes that appeared in the original design, and that were used to ensure spatial distribution of the sampled lakes over the region of interest. Larsen *et al.* (1993) provide details on the construction of the spatial clusters.

Let H be the number of strata, n_h the number of observations within stratum h , and A_h the set of sampled elements that fall in stratum h . Define $p_k = n_h^{-1} \pi_k$. Using this notation and the assumption of a stratified sample with replacement, we rewrite the semiparametric estimator as

$$\hat{y}_{\text{reg}} = \frac{1}{N} \sum_{k \in U} \hat{g}(x_k, z_k) + \frac{1}{N} \sum_{h \in H} \frac{1}{n_h} \sum_{k \in A_h} \frac{y_k - \hat{g}(x_k, z_k)}{p_k} \quad (12)$$

and the variance estimator as

$$\hat{V}(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum_{h \in H} S_h^2,$$

where S_h^2 is the estimated within-stratum weighted residual variance for stratum h . Assuming the strata are sampled with replacement, Särndal *et al.* (1992, page 421-422) suggest S_h^2 can be calculated as

$$S_h^2 = \frac{1}{n_h(n_h - 1)} \sum_{k \in A_h} \left(\frac{y_k - \hat{g}(x_k, z_k)}{p_k} - \frac{\sum_{l \in A_h} \frac{y_l - \hat{g}(x_l, z_l)}{\pi_l}}{\pi_l} \right)^2. \quad (13)$$

Similarly, we estimate $\hat{V}(\hat{y}_H)$ through

$$\hat{V}(\hat{y}_H) = \frac{1}{\hat{N}^2} \sum_{h \in H} \frac{1}{n_h(n_h - 1)} \sum_{k \in A_h} \left(\frac{p_k}{-\sum_{l \in A_h} \frac{y_l - \hat{y}_H}{\pi_l}} \right)^2, \quad (14)$$

and the expression for $\hat{V}(\hat{y}_{\text{par}})$ is obtained completely analogously as for $\hat{V}(\hat{y}_{\text{reg}})$ except that $\hat{g}(x_k, z_k)$ is computed by linear regression.

This setup allows us to obtain the following estimates of mean ANC for the Northeastern lakes, together with variance estimates and approximate 95% confidence intervals (CI). A local linear fit has been employed for the nonparametric term with bandwidth set at one tenth of the range of UTMX.

$$\hat{y}_{\text{reg}} = 558.0 \text{ } \mu\text{eq/L} \quad \hat{V}(\hat{y}_{\text{reg}}) = 2534.6 \quad \text{CI} = (459.3; 656.6)$$

$$\hat{y}_{\text{par}} = 577.3 \text{ } \mu\text{eq/L} \quad \hat{V}(\hat{y}_{\text{par}}) = 3239.6 \quad \text{CI} = (465.8; 688.9)$$

$$\hat{y}_H = 555.9 \text{ } \mu\text{eq/L} \quad \hat{V}(\hat{y}_H) = 4313.3 \quad \text{CI} = (427.2; 684.7)$$

The confidence interval constructed using the Hájek estimator is about 31% wider than that constructed using the semiparametric estimator, while the interval for the fully parametric regression estimator is 13% wider. These results show evidence of an improvement in efficiency provided by accounting for the auxiliary information in both a parametric and nonparametric way in the mean estimation procedure, with the nonparametric estimator able to capture some additional efficiency beyond that of the parametric estimator.

As mentioned above, an important goal of this application is the assessment of how many lakes are at risk of acidification or are acidified already. That is, we are interested in estimating the proportion of Northeastern lakes with ANC values smaller than some specific threshold values. We can determine such proportions by estimating the finite population distribution function,

$$F_N(t) = \frac{1}{N} \sum_{k \in U} I_{\{y_k \leq t\}}$$

at specific threshold values t , where $I_{\{y_k \leq t\}}$ denotes the indicator function taking a value of 1 if $y_k \leq t$ and 0 otherwise. Because all three estimators can be expressed as weighted sums of sample observations, the weights obtained for each can be applied directly to the $I_{\{y_k \leq t\}}$ for the sample to estimate $F_N(t)$ for any desired t . Let us denote by $\hat{F}_H(t)$, $\hat{F}_{\text{reg}}(t)$ and $\hat{F}_{\text{par}}(t)$ the Hájek, semiparametric and

parametric regression estimators of the distribution function, respectively. Estimates for their design variances are computed by plugging the indicator variables in equations (13) and (14).

Figure 1 shows estimates of the ANC cdf produced by $\hat{F}_H(t)$, $\hat{F}_{\text{par}}(t)$ and $\hat{F}_{\text{reg}}(t)$ evaluated on a grid of 1,000 equally spaced values for t . Included are their respective pointwise 95% confidence intervals calculated at each grid point. All three estimators are similar, but the confidence bands for the parametric and semiparametric regression estimators tend to be narrower. Averaged over all 1,000 grid points, the widths of the confidence bands are 0.093 for $\hat{F}_H(t)$, 0.084 for $\hat{F}_{\text{par}}(t)$ and 0.075 for $\hat{F}_{\text{reg}}(t)$, respectively.

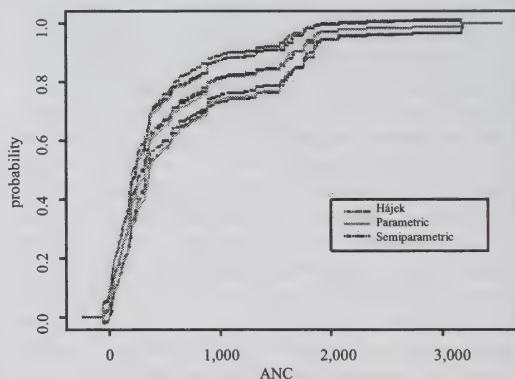


Figure 1
Estimates of the population cumulative distribution function for ANC and confidence bounds produced by Hájek, parametric and semiparametric regression estimators

Along with ANC, the EMAP survey of Northeastern lakes measured the concentration of multiple chemistry variables including sulfate, magnesium and chloride, so that the survey weights obtained for ANC can also be applied to these concentrations as well as their respective cdfs. As another illustration of the semiparametric estimation approach, it is possible to “invert” $\hat{F}_{\text{reg}}(t)$ to obtain quantile estimators $\hat{\theta}_{\text{reg}}(\alpha) = \min\{t: \hat{F}_{\text{reg}}(t) \geq \alpha\}$ of these additional chemistry variables. Table 1 displays semiparametric estimates of the first, second, and third quartiles of sulfate, magnesium, and chloride measured in ($\mu\text{eq/L}$). Variance estimation for these quantiles could be handled using asymptotic results of Francisco and Fuller (1991), but will not be explored further here.

Table 1 Quartile estimates of chemistry variables

α	Sulfate	Magnesium	Chloride
0.25	73.3	63.8	27.4
0.50	104.3	127.0	162.2
0.75	201.4	221.9	462.2

5. Conclusion

In this article, we have described a model-assisted estimator that uses semiparametric regression to capture relationships between multiple population-level auxiliary variables and the survey variables. We have developed asymptotic theory that shows the resulting estimator is design consistent and asymptotically normal under mild conditions on the design and the population. This generalizes the results of Breidt and Opsomer (2000), who had proved similar results for a univariate nonparametric model-assisted estimator. The semiparametric estimator was applied to data from a survey of lakes in the Northeastern U.S., where it was shown to be more efficient than an estimator that does not take advantage of the auxiliary variables and than a fully parametric regression estimator.

In addition to its theoretical properties, the semiparametric model-assisted estimator has attractive practical properties as well. As noted earlier, it is fully calibrated for the auxiliary variables, whether used in the parametric or nonparametric model components, and it is location and scale invariant. The estimator can be expressed as a weighted sum of the sample observations, so that it conforms to the traditional survey estimation paradigm and a single set of weights can be applied to all the survey variables, hence preserving relationships between variables.

One issue which was not addressed in the current article is the selection of the smoothing parameter for the nonparametric component of the regression model. This is a challenging topic in the model-assisted context, further complicated by the just mentioned fact that a single set of survey regression weights is applied to all the survey variables: because the optimal bandwidth choice depends on the variable being smoothed, no single bandwidth (and hence set of weights) will be optimal for all variables in the survey. This topic is currently being explored by the authors.

Acknowledgments

The research for this article was supported by National Science Foundation grants DMS-0204531 and DMS-0204642, and by STAR Research Assistance Agreements CR-829095 and CR-829096 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University and Oregon State University. This manuscript has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

Appendix

Technical assumptions and derivations

We begin by stating the necessary assumptions, which extend those used in Breidt and Opsomer (2000) to the semiparametric model.

Assumptions:

- **A1** *Distribution of the errors under ξ : the errors ε_k are independent and have mean zero, variance $v(x_k, z_k)$, and compact support, uniformly for all N .*
- **A2** *Distribution of the covariates: the x_k and z_k are considered fixed with respect to the superpopulation model ξ . The z_k are assumed to have bounded support, and the x_k are independent and identically distributed $F(x) = \int_{-\infty}^x f(t)dt$, where $f(\cdot)$ is a density with compact support $[a_x, b_x]$ and $f(x) > 0$ for all $x \in [a_x, b_x]$.*
- **A3** *Nonparametric mean and variance functions: the mean function $m(\cdot)$ is continuous, and the variance function $v(\cdot, \cdot)$ is bounded and strictly greater than 0.*
- **A4** *Kernel K : the kernel $K(\cdot)$ has compact support $[-1, 1]$, is symmetric and continuous, and satisfies $\int_{-1}^1 K(u)du = 1$.*
- **A5** *Sampling rate nN^{-1} and bandwidth h_N : as $N \rightarrow \infty$, $nN^{-1} \rightarrow \pi \in (0, 1)$, $h_N \rightarrow 0$ and $Nh_N^2 / (\log \log N) \rightarrow \infty$.*
- **A6** *Inclusion probabilities π_k and π_{kl} : for all N , $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k, l \in U_N} \pi_{kl} \geq \lambda^* > 0$ and*

$$\limsup_{N \rightarrow \infty} \max_{k, l \in U_N, k \neq l} |\pi_{kl} - \pi_k \pi_l| < \infty.$$
- **A7** *Additional assumptions involving higher-order inclusion probabilities:*

$$\lim_{N \rightarrow \infty} n^2 \max_{(k_1, k_2, k_3) \in D_{3, N}} |E_p(I_{k_1} - \pi_{k_1})(I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})(I_{k_4} - \pi_{k_4})| < \infty,$$
where $D_{t, N}$ denotes the set of all distinct t -tuples (k_1, k_2, \dots, k_t) from U_N ,

$$\lim_{N \rightarrow \infty} \max_{(k_1, k_2, k_3, k_4) \in D_{4, N}} |E_p(I_{k_1} I_{k_2} - \pi_{k_1 k_2})(I_{k_3} I_{k_4} - \pi_{k_3 k_4})| = 0,$$
and

$$\limsup_{N \rightarrow \infty} \max_{(k_1, k_2, k_3) \in D_{3, N}} |E_p(I_{k_1} - \pi_{k_1})^2 (I_{k_2} - \pi_{k_2})(I_{k_3} - \pi_{k_3})| < \infty.$$
- **A8** *The matrix $N^{-1} \mathbf{Z}_U^T (\mathbf{I} - \mathbf{S}_U^*) \mathbf{Z}_U$ is invertible for all N with model probability 1.*

Assumption A8 is required so that the population estimator \mathbf{B} is well-defined. The invertibility of the matrix in A8 depends on the combined effect of the bandwidth h and the joint distribution of the x_k and z_k . While it would in principle be possible to write down sufficient conditions for this, we opted for this simpler and more explicit approach.

Before giving the proofs of Theorems 3.1 and 3.2, we state and prove a number of lemmas.

Lemma 1 Under the assumptions A1-A7,

(a) for all $k \in U$ and $d = 1, \dots, D$,

$$\frac{1}{N} \sum_U E_p (s_{Ak}^T \mathbf{Y}_A - s_{Uk}^T \mathbf{Y}_U)^2 = O\left(\frac{1}{nh}\right)$$

and

$$\frac{1}{N} \sum_U E_p (s_{Ak}^T \mathbf{Z}_{dA} - s_{Uk}^T \mathbf{Z}_{dU})^2 = O\left(\frac{1}{nh}\right);$$

(b) the $s_{Uk}^T \mathbf{Y}_U$ and $s_{Uk}^T \mathbf{Z}_U$ are uniformly bounded over all $k \in U$.

Proof of Lemma 1: Since both the y_k and z_{dk} are bounded by assumption, part (a) can be shown using an identical reasoning as in Lemma 4 of Breidt and Opsomer (2000). While that lemma did not include a rate of convergence, that rate is readily derived by noting that

$$\frac{1}{N} \sum_{k \in U} z_{ik}^2 = O\left(\frac{1}{nh}\right)$$

in the notation of Breidt and Opsomer (2000) and then proceeding as in that proof.

Part (b) was proven directly in Lemma 2 (iv) of Breidt and Opsomer (2000).

Lemma 2 Under assumptions A1-A8,

$$\hat{\mathbf{B}} = \mathbf{B} + O_p(1/\sqrt{nh}),$$

with the rate holding component-wise, and \mathbf{B} is bounded for all N .

Proof of Lemma 2: Write $\tilde{y}_k^{[s_U]} = s_{Uk}^T \mathbf{Y}_U$ and $\tilde{y}_k^{[s_A]} = s_{Ak}^T \mathbf{Y}_A$ for the population and sample smoothed versions of y_k , and similarly, $\tilde{z}_k^{[s_U]} = s_{Uk}^T \mathbf{Z}_U$ and $\tilde{z}_k^{[s_A]} = s_{Ak}^T \mathbf{Z}_A$. We rewrite expression (6) as a function of sample-weighted terms \hat{t}_l , $l = 1, \dots, 6$:

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{t}_1 & \hat{t}_2 \\ \hat{t}_3 & \hat{t}_4 \end{bmatrix}^{-1} \begin{bmatrix} \hat{t}_5 \\ \hat{t}_6 \end{bmatrix},$$

where

$$\hat{t}_1 = \left(\frac{\hat{N}}{N}\right)^2$$

$$\hat{t}_2 = \bar{z}_\pi - \frac{1}{N} \sum_A \frac{\tilde{z}_k^{[s_A]}}{\pi_k} \left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{t}_3 = \bar{z}_\pi^T \left(\frac{\hat{N}}{N}\right)$$

$$\hat{t}_4 = \frac{1}{N} \sum_A \frac{\mathbf{z}_k^T \mathbf{z}_k}{\pi_k} - \frac{1}{N} \sum_A \frac{\mathbf{z}_k^T \tilde{\mathbf{z}}_k^{[s_A]}}{\pi_k} + \bar{z}_\pi^T \frac{1}{N} \sum_A \frac{\tilde{\mathbf{z}}_k^{[s_A]}}{\pi_k}$$

$$\hat{t}_5 = \bar{y}_\pi - \frac{1}{N} \sum_A \frac{\tilde{y}_k^{[s_A]}}{\pi_k} \left(1 - \frac{\hat{N}}{N}\right)$$

$$\hat{t}_6 = \frac{1}{N} \sum_A \frac{\mathbf{z}_k^T \mathbf{y}_k}{\pi_k} - \frac{1}{N} \sum_A \frac{\mathbf{z}_k^T \tilde{\mathbf{y}}_k^{[s_A]}}{\pi_k} + \bar{z}_\pi^T \frac{1}{N} \sum_A \frac{\tilde{\mathbf{y}}_k^{[s_A]}}{\pi_k}.$$

The sample-weighted estimator $\hat{\mathbf{B}}$ will be expanded around

$$\mathbf{B} = \begin{bmatrix} 1 & \bar{z}_\pi \\ \bar{z}_\pi^T & \mathbf{t}_4 \end{bmatrix}^{-1} \begin{bmatrix} \bar{y}_\pi \\ \mathbf{t}_6 \end{bmatrix}, \quad (15)$$

where

$$\mathbf{t}_4 = \frac{1}{N} \sum_U \mathbf{z}_k^T \mathbf{z}_k - \frac{1}{N} \sum_U \mathbf{z}_k^T \tilde{\mathbf{z}}_k^{[s_U]} + \bar{z}_\pi^T \frac{1}{N} \sum_U \tilde{\mathbf{z}}_k^{[s_U]}$$

$$\mathbf{t}_6 = \frac{1}{N} \sum_U \mathbf{z}_k^T \mathbf{y}_k - \frac{1}{N} \sum_U \mathbf{z}_k^T \tilde{\mathbf{y}}_k^{[s_U]} + \bar{z}_\pi^T \frac{1}{N} \sum_U \tilde{\mathbf{y}}_k^{[s_U]}$$

and the remaining \mathbf{t}_l can be found in (15). The existence and continuity of the derivatives of $\hat{\mathbf{B}}$ with respect to the \hat{t}_l and evaluated at \mathbf{t}_l follow from Lemma 1(b) and the existence of the inverse in (15), which is assumed by A8.

The result will follow from a 0th order Taylor expansion if we can show that $\hat{t}_l - \mathbf{t}_l = O_p(1/\sqrt{nh})$ for all l (e.g., Fuller (1996), Corollary 5.1.5). For \hat{t}_1 and \hat{t}_3 , this follows directly from A2 and A6. The remaining terms contain sums involving smoothed quantities $\tilde{\mathbf{z}}_k^{[s_A]}$ and $\tilde{\mathbf{y}}_k^{[s_A]}$. We demonstrate the reasoning for one of those terms in \hat{t}_6 . We have

$$\begin{aligned} \frac{1}{N} \sum_A \frac{\mathbf{z}_k^T \tilde{\mathbf{y}}_k^{[s_A]}}{\pi_k} - \frac{1}{N} \sum_U \mathbf{z}_k^T \tilde{\mathbf{y}}_k^{[s_U]} &= \frac{1}{N} \sum_U \mathbf{z}_k^T \tilde{\mathbf{y}}_k^{[s_U]} \left(\frac{\mathbf{I}_k}{\pi_k} - 1 \right) \\ &\quad + \frac{1}{N} \sum_U \mathbf{z}_k^T (\tilde{\mathbf{y}}_k^{[s_A]} - \tilde{\mathbf{y}}_k^{[s_U]}) \frac{\mathbf{I}_k}{\pi_k}, \end{aligned}$$

and the first term is $O_p(1/\sqrt{n})$ by A6 and Lemma 1(b), using the same argument as in Lemma 4 of Breidt and Opsomer (2000). For the second term, use Schwarz's inequality

$$\begin{aligned} &\left| \frac{1}{N} \sum_U \mathbf{z}_k^T (\tilde{\mathbf{y}}_k^{[s_A]} - \tilde{\mathbf{y}}_k^{[s_U]}) \frac{\mathbf{I}_k}{\pi_k} \right| \\ &\leq \sqrt{\frac{1}{N} \sum_U \mathbf{z}_k^{[2T]} \frac{\mathbf{I}_k}{\pi_k^2}} \sqrt{\frac{1}{N} \sum_U (\tilde{\mathbf{y}}_k^{[s_A]} - \tilde{\mathbf{y}}_k^{[s_U]})^2}, \end{aligned}$$

where $z_k^{[2]}$ denotes that the squares are computed component-wise. The first term is bounded by A2 and A6, and the second term is $O_p(1/\sqrt{nh})$ by Lemma 1(a) and Markov's inequality. The desired result then follows by applying the same reasoning to the remaining terms in $\hat{z}_2, \hat{z}_4, \hat{z}_5, \hat{z}_6$.

The boundedness of \mathbf{B} follows directly from assumption A8, Lemma 1(b) and the boundedness of the z_k .

Lemma 3 Under the assumptions A1-A8, we have

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof of Lemma 3: Given expression (9), we need to show that

$$(\bar{z}_N - \bar{z}_n)(\mathbf{B} - \hat{\mathbf{B}}) = o_p\left(\frac{1}{\sqrt{n}}\right) \quad (16)$$

$$\frac{1}{N} \sum_U (m_k - \hat{m}_k) \left(1 - \frac{I_k}{\pi_k}\right) = o_p\left(\frac{1}{\sqrt{n}}\right). \quad (17)$$

Lemma 2 and assumptions A2, A5 and A6 show that $(\bar{z}_N - \bar{z}_n)(\mathbf{B} - \hat{\mathbf{B}}) = O_p(1/nh)$. In order to prove (17), we can rewrite it as

$$\begin{aligned} \frac{1}{N} \sum_U (m_k - \hat{m}_k) \left(1 - \frac{I_k}{\pi_k}\right) &= \frac{1}{N} \sum_U (\bar{y}_k^{[s_U]} - \bar{y}_k^{[s_A]}) \left(1 - \frac{I_k}{\pi_k}\right) \\ &\quad - \frac{1}{N} \sum_U (\bar{z}_k^{[s_U]} - \bar{z}_k^{[s_A]}) \left(1 - \frac{I_k}{\pi_k}\right) \mathbf{B} \\ &\quad - \frac{1}{N} \sum_U \bar{z}_k^{[s_A]} \left(1 - \frac{I_k}{\pi_k}\right) (\mathbf{B} - \hat{\mathbf{B}}). \end{aligned}$$

The first term on the right hand side has been proven to be $O_p(1/\sqrt{n})$ in Lemma 5 of Breidt and Opsomer (2000); this same Lemma and boundness of \mathbf{B} provide the same rate for the second term. Assumptions A5-A6, Lemma 1(b) and Lemma 2 show that the third term is $O_p(1/n\sqrt{h})$ and the desired rate is achieved.

Lemma 4 Under assumptions A6 and A8,

$$\begin{aligned} E_p(\hat{y}_{\text{dif}}) &= \bar{y}_N \\ \text{Var}_p(\hat{y}_{\text{dif}}) &= \frac{1}{N^2} \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k - g_k}{\pi_k} \frac{y_l - g_l}{\pi_l} \\ &= O\left(\frac{1}{n}\right). \end{aligned}$$

Proof of Lemma 4: The properties of the difference estimator are readily computed. The rate of the design variance follows from the stated assumptions using the same reasoning as in Lemma 4 of Breidt and Opsomer (2000).

Lemma 5 Under assumptions A1-A8,

$$\hat{V}(\hat{y}_{\text{reg}}) = \text{Var}_p(\hat{y}_{\text{dif}}) + o_p\left(\frac{1}{n}\right).$$

Proof of Lemma 5: The reasoning for this proof will closely follow that of Theorem 3 of Breidt and Opsomer (2000). We write

$$\begin{aligned} \hat{V}(\hat{y}_{\text{reg}}) - \text{Var}_p(\hat{y}_{\text{dif}}) &= (\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) \\ &\quad + (\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})) \quad (18) \end{aligned}$$

with

$$\hat{V}(\hat{y}_{\text{dif}}) = \frac{1}{N^2} \sum_A \sum_A \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_k - g_k}{\pi_k} \frac{y_l - g_l}{\pi_l}.$$

Since

$$\frac{1}{N} \sum_U (y_k - g_k)^4 < \infty,$$

by assumptions A1-A3 and from Lemmas 1(b) and 2, the approach used for the term A_N of Breidt and Opsomer (2000) can be used to show that

$$E_p |\hat{V}(\hat{y}_{\text{dif}}) - \text{Var}_p(\hat{y}_{\text{dif}})| = o\left(\frac{1}{n}\right),$$

which provides the desired consistency by the Markov inequality.

For the first term in (18), note that

$$\begin{aligned} \hat{g}_k - g_k &= (\bar{y}_k^{[s_A]} - \bar{y}_k^{[s_U]}) - (\bar{z}_k^{[s_A]} - \bar{z}_k^{[s_U]}) (\hat{\mathbf{B}} - \mathbf{B}) \\ &\quad + (\bar{z}_k - \bar{z}_k^{[s_U]}) (\hat{\mathbf{B}} - \mathbf{B}) - (\bar{z}_k^{[s_A]} - \bar{z}_k^{[s_U]}) \mathbf{B}, \end{aligned}$$

so that

$$\begin{aligned} (\hat{V}(\hat{y}_{\text{reg}}) - \hat{V}(\hat{y}_{\text{dif}})) &= \\ \frac{1}{N^2} \sum_U \sum_U &\left\{ \frac{-2 \frac{y_k - g_k}{\pi_k} \frac{\hat{g}_l - g_l}{\pi_l}}{\frac{\hat{g}_k - g_k}{\pi_k} \frac{\hat{g}_l - g_l}{\pi_l}} \right\} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} I_k I_l \end{aligned}$$

can be decomposed into variance terms involving sample and population smooths and parameter estimators. Each of these terms can be shown to be $O_p(1/n)$. We demonstrate the approach on one of the terms:

$$\begin{aligned} &\left| \frac{1}{N^2} \sum_U \sum_U \frac{y_k - g_k}{\pi_k} \frac{\hat{z}_l - z_l}{\pi_l} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} I_k I_l (\hat{\mathbf{B}} - \mathbf{B}) \right| \\ &\leq \left(\frac{C_1}{N} + C_2 \max |\pi_{kl} - \pi_k \pi_l| \right) \frac{1}{N} \sum_U |\bar{z}_k^{[s_A]} - \bar{z}_k^{[s_U]}| \|\hat{\mathbf{B}} - \mathbf{B}\| \\ &= o_p\left(\frac{1}{n}\right) \end{aligned}$$

where $C_1, C_2 < \infty$ summarize the bounded terms (by assumptions A1-A3 and A6 and Lemma 1(b)), and the rate of convergence is the result of assumption A6 and Lemmas 1(a) and 2.

Proof of Theorem 3.1: In Lemma 3, we show that

$$\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where \hat{y}_{dif} is the difference estimator (3). The result immediately follows from assumption A5 and Lemma 4.

Proof of Theorem 3.2: Note that \hat{y}_{dif} can be written as the sum of a population constant and an expansion estimator of the form \bar{y}_π by defining a new variable $y_k - s_{Uk}^T Y_U + s_{Uk}^T Z_U B - z_k B$ for $k \in U$. As is the case for the original y_k , this new variable has bounded support by Lemma 1(b) and a variance of order $O(1/n)$ by Lemma 4. Hence, existence of the CLT for \bar{y}_π implies existence of the CLT for \hat{y}_{dif} . Also, $\hat{y}_{\text{reg}} = \hat{y}_{\text{dif}} + o_p(1/\sqrt{n})$ by Lemma 3, so that $\sqrt{n} \hat{y}_{\text{reg}}$ and $\sqrt{n} \hat{y}_{\text{dif}}$ have the same asymptotic distribution. Applying Slutsky's Theorem and Lemma 5 complete the proof.

References

- Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2nd Ed.). New York: John Wiley & Sons, Inc.
- Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models*. Washington, D.C.: Chapman and Hall.
- Isaki, C., and Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Larsen, D.P., Thornton, K.W., Urquhart, N.S. and Paulsen, S.G. (1993). Overview of survey design and lake selection. EMAP - Surface Waters 1991 Pilot Report. Technical Report EPA/620/R-93/003, U.S. Environmental Protection Agency. (Eds. D.P. Larsen and S.J. Christie).
- Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.
- Opsomer, J.D., Breidt, F.J., Moisen, G.G. and Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*. To appear.
- Opsomer, J.-D., and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.
- Opsomer, J.D., and Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, 8, 715-732.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Speckman, P.E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society, Series B*, 50, 413-436.
- Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Webb, J.R. and Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Technical Report EPA/620/R-03/001, U.S. Environmental Protection Agency, Washington, DC.
- U.S. National Acid Precitation Assessment Program (1991, November). 1990 Integrated Assessment Report. Technical report, Washington, DC.

Ex post weighting of price data to estimate depreciation rates

Marc Tanguay and Pierre Lavallée¹

Abstract

To model economic depreciation, a database is used that contains information on assets discarded by companies. The acquisition and resale prices are known along with the length of use of these assets. However, the assets for which prices are known are only those that were involved in a transaction. While an asset depreciates on a continuous basis during its service life, the value of the asset is only known when there has been a transaction. This article proposes an *ex post* weighting to offset the effect of source of error in building econometric models.

Key Words: Price ratio; Survival data; Uniform distribution; Depreciation of vehicles.

1. Introduction

Various econometric models are used to estimate economic depreciation. To this end, we use a database containing information on assets discarded by companies. The acquisition and resale prices are known along with the length of use of these assets. From this information, we would like to infer results for the total population of assets used by companies. Regarding the use of the prices of used assets to estimate economic depreciation, we refer the reader to, Gellatly, Tanguay and Yan (2002) and Hulten and Wykoff (1981).

We question, however, the representativeness of the database used. Indeed, the assets for which prices are known are solely those subject to a transaction. We do not know the extent to which the losses of value observed on these assets are representative of the loss of value for all assets in production, regardless of whether they were the subject of a transaction. This situation can be a source of error in building econometric models because these models seek to measure depreciation of assets over their service lives, regardless of whether there was a transaction.

It is this second source of error that we propose to offset, at least in part, by applying *ex post* weighting when building econometric models. Section 2 of this article will describe the problem in greater detail, while in Section 3, we will describe the approach used to determine the weights. Finally, in Section 4, we present some numeric results.

2. Problem

We are seeking to describe the relationship between prices and asset age. There is a sample of n assets where we know, for each asset i , the price ratio r_i and the time t_i when this ratio was measured. Once prices are expressed in

real dollars, this ratio is given as $r_i = P_i' / P_i^0$ where P_i^0 is the initial value of the investment in asset i and P_i' is its resale price at time t . This ratio is strictly decreasing in relation to the time axis t . At the start, we do not know the process that generates the loss in value and there are no specifics about the function that describes this loss except that it is strictly decreasing. However, it is possible to examine the distribution of the price ratios between 0 and 1. Here is an example constructed from data on manufacturing plants (note that 2/3 of the sample was excluded because it corresponds to discarded assets (the price is zero) and the estimation procedures take this component into account, each in its own way).

Since we want to use the data to infer statistics on the population of assets in production, we would like our data to have properties similar to those of a random sample drawn from that population. As we stated earlier, this is not the case because we only have the prices of assets i that were subject to a transaction at time t_i , $i = 1, \dots, n$. In effect, while we would like to have price ratios for various periods in the existence of a given asset i , the ratio is only available when there has been a transaction, something that occurs in a non-uniform manner over an asset's service life.

Consequently, we can ask ourselves what form the above distribution might have if it had been drawn from a sample in which the price ratio had been measured, for the same asset i at different times t . Our argument is that it should converge toward a *uniform distribution*. We will therefore seek to obtain a weighting that will help us recreate a uniform distribution of price ratios. This weighting will help us offset the lack of uniformity in the distribution of observations, which may impact statistical analyses such as linear regression.

1. Marc Tanguay and Pierre Lavallée, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. E-mail: marc.tanguay@statcan.ca, pierre.lavallee@statcan.ca.

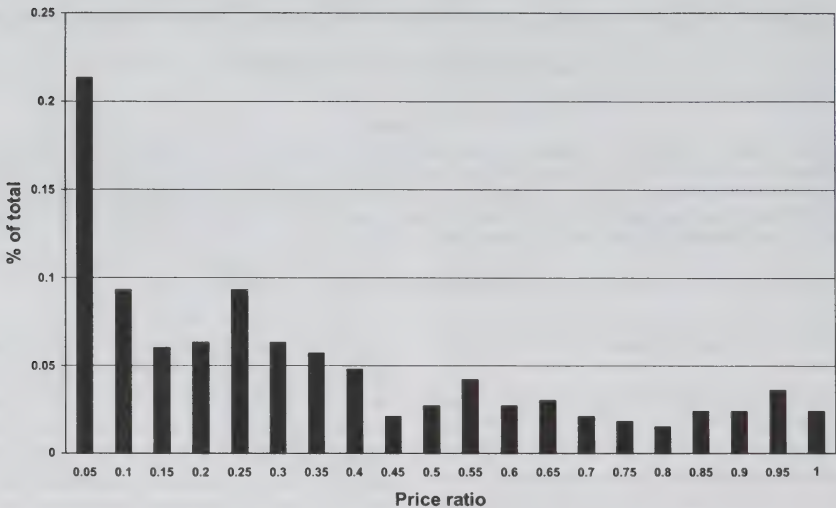


Figure 1 Distribution of observations by price ratio, manufacturing plants

3. Approach

Our starting point is that price ratios can be considered empirical realizations of an unknown form of survival function. In service life models, the survival function expresses the probability that an entity with a limited service life will survive beyond a certain point on the time axis. Accordingly, it provides the same information as a distribution function (or Cumulative Distribution Function). We will let r_t be a random variable describing the service life of a unit of value incorporated in some asset. The value gradually erodes over time for as long as the asset is in service. The price ratio can therefore be interpreted as the surviving fraction that gradually becomes smaller and smaller. This fraction is written as $S(y)$ and gives

$$S(y) = 1 - F(y)$$

where $F(y) = P(r \leq y)$ is the distribution function, that is, the probability that a unit of value is lost before point y .

Fundamental transformation theorems of probability laws provide the means for defining the inverse function of $F(y)$ (Greene 1993 and Ross 2002). We let $z = F(y)$ and assume that the inverse function F^{-1} exists so that $y = F^{-1}(z)$. This shows that there is a direct match between the space of y , bounded at 0 but infinite to the right, and that of F which is bound between 0 and 1. The distribution function of z is $F(F^{-1}(z)) = z$. The law that generates this distribution is a uniform distribution between 0 and 1.

This result is generally at the core of data generation processes like Monte Carlo simulations because the uniform distribution is often used when a random sample is being generated, followed by the application of the inverse function (Davidson and MacKinnon 1993). This approach is not always practical and indeed is sometimes patently impossible, especially if the inverse function, F^{-1} is not explicit. This result has also been used in generalized remainder approaches, notably to build specification tests (Lancaster 1985).

The result is that any random sample built using empirical realizations of survival proportion data must converge in distribution toward a uniform distribution.

In the case of price data, intuition suggests that between the time of investment and that of disposal, the full range of relative prices must be covered by an asset in production. Initially, value depreciates faster and therefore there are more observations with short periods of time. This is offset by the fact that the corresponding reference on the time scale is also shorter. For example, it takes less time to move from 100% of the initial value to 90%, than from 15% to 5% of the initial value.

It is easy to verify these findings numerically using simulated data and we will not spend time on this. Rather, we will examine how this result can be reintroduced in the database to produce, at least partially, properties similar to those of a random sample. *We can do this by simply imposing ex post on the empirical price distribution a*

weight structure w_i that ensures that the empirical distribution of the data, in the price space, is uniform.

The empirical distribution of price ratios r is given by

$$\hat{F}_n(y) = \frac{\sum_{i=1}^n I_i(y)}{n} \quad (1)$$

where $I_i(y) = 1$ if the measured value r_i of asset i is less than or equal to y (specifically, $r_i \leq y$), and 0 otherwise, and n is the total number of observations. Note that if the n units of the sample are independent and identically distributed (i.i.d.), when $n \rightarrow \infty$, $\hat{F}_n(y)$ converges in probability to $F(y)$, that is, $\hat{F}_n(y) \xrightarrow{P} F(y)$ (Bickel and Doksum 1977).

To obtain weight w_i for each asset i , we simply distribute the sample in a given number H of intervals (or classes) of a fixed size on the scale of price ratios, and we assign the same probability $\pi = 1/H$ to each of these intervals. Since the price ratios are bounded by 0 and 1, we then have the interval $h=1$ given by $[0, H^{-1}]$, and for $h=2, \dots, H$, the intervals are given by $[(h-1)H^{-1}, hH^{-1}]$. A weight w_h is then calculated in each interval h by the ratio $\pi/\hat{\pi}_h$ where $\hat{\pi}_h$ is the empirical probability specific to interval h , producing

$$\hat{\pi}_h = \frac{1}{n} \sum_{i=1}^n \delta_i(h) = \frac{n_h}{n} \quad (2)$$

where $\delta_i(h) = 1$ if $r_i \in h$, 0 otherwise. We then propose

$$\begin{aligned} w_i = w_h &= \frac{\pi}{\hat{\pi}_h} \\ &= \frac{n}{Hn_h} \end{aligned} \quad (3)$$

for $r_i \in h$. Using these weights, the weighted empirical distribution of the price ratios r is given by

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^n w_i I_i(y)}{\sum_{i=1}^n w_i}. \quad (4)$$

By writing $\sum_{i=1}^n w_i = \sum_{h=1}^H \sum_{i=1}^{n_h} n/Hn_h = n$, we finally get

$$\hat{F}_{n,w}(y) = \frac{\sum_{i=1}^n w_i I_i(y)}{n}. \quad (5)$$

Since $n_h = \sum_{i=1}^n \delta_i(h)$, we have

$$\begin{aligned} \hat{F}_{n,w}(y) &= \frac{\sum_{i=1}^n w_i I_i(y)}{n} \\ &= \frac{1}{H} \sum_{h=1}^H \frac{1}{n_h} \sum_{i=1}^n \delta_i(h) I_i(y) \\ &= \frac{1}{H} \sum_{h=1}^H \frac{\sum_{i=1}^n \delta_i(h) I_i(y)}{\sum_{i=1}^n \delta_i(h)} \\ &= \frac{1}{H} \sum_{h=1}^H \hat{F}_n(y|h). \end{aligned} \quad (6)$$

When $n \rightarrow \infty$, we have $(1/n) \sum_{i=1}^n \delta_i(h) I_i(y) \xrightarrow{P} P(r \in h, r \leq y)$ and $(1/n) \sum_{i=1}^n \delta_i(h) \xrightarrow{P} P(r \in h)$. Thus, when $n \rightarrow \infty$,

$$\begin{aligned} \hat{F}_n(y|h) &\xrightarrow{P} \frac{P(r \in h, r \leq y)}{P(r \in h)} \\ &= P(r \leq y | r \in h) = F(y|h) \end{aligned} \quad (7)$$

where $F(y|h)$ is the distribution of price ratios r within interval h .

For a sufficiently large n , H must be determined in such a way as to build the intervals h so that $\hat{F}_n(y|h)$ is distributed approximately uniformly, $h=1, \dots, H$. In other words, when $n \rightarrow \infty$, for a sufficiently large H , $F(y|h)$ should have a uniform distribution on interval h . Note that this argument was used by Dalenius and Hodges (1959) in a context of optimal stratification. In this case, the distribution $F(y|h)$ is given by

$$F(y|h) = \begin{cases} 0 & \text{for } y \leq (h-1)H^{-1} \\ Hy - h + 1 & \text{for } (h-1)H^{-1} < y \leq hH^{-1} \\ 1 & \text{for } y > hH^{-1} \end{cases} \quad (8)$$

Since $F(y) = \sum_{h=1}^H F(y|h)/H$, we have $F(y) = y$, which corresponds to the uniform distribution. We conclude from this that for a sufficiently large n , the use of weighting (3) should ensure that the weighted empirical distribution $\hat{F}_{n,w}(y)$ given by (5) is distributed approximately uniformly.

Monte Carlo simulations have shown that estimates produced from a non-random sample could be improved by using this approach. Its main advantages can be attributed to:

- its simplicity;
- the fact that it can be introduced *ex ante*, or prior to introducing the econometric model as such. Consequently, it does not require strong working hypotheses.

If we go back to the histogram presented earlier and divide the sample in $H = 5$ intervals of a width of 0.2 and a value of $\pi = 1/5 = 0.2$, we then get the following histogram that was weighted *ex post*.

4. Application

We will now illustrate our approach using an example taken from the Kelly Blue Book, a source of information widely used to estimate depreciation of automobiles. Table 1 shows the prices of two models of cars at different ages between 1 and 18 years. For each car, we have a sample of $n = 18$ units. Prices are expressed in relative value in

relation to a new model. The ratios also have to be adjusted to take into account the survival probability at each of these ages. For each vehicle, the final ratio used r_i for year i is built from the product of the price ratio times the survival probability.

We are interested in the average depreciation rate $\bar{\tau}$ for each car. This can be estimated from a regression of the prices (or from a function of these prices) in relation to age (or a function of age). However, if we assume that the rate is constant and geometric, we obtain the relationship $r_i = 1 - \bar{\tau}^i$, where r_i is the relative price based on age i . In this case, a rate $\hat{\tau}_i$ can be estimated at each age i by $\hat{\tau}_i = 1 - r_i^{1/i}$. An estimate of the average rate of depreciation is then produced from the average for all ages, $\hat{\tau} = \sum_{i=1}^{18} \hat{\tau}_i / 18$.

In the above example, we see that the depreciation rates $\hat{\tau}_i$ vary by age range and that they tend to increase with age. Moreover, the fact that we use a simple average of the ages in calculating $\hat{\tau}$ again implicitly gives the same weight to each age. However, it is quite clear that this is not the distribution that we would get from a random sample of service vehicles. The figure below shows the distribution of price cells between ratios of 0 and 1.

The reweighting technique simply involves applying an equal weight to each of the relative price ranges. In this example, the $n=18$ ages are distributed into $H=7$ classes, resulting in 18/7 of the ages in each class (in reality, the structures of the cells was configured into 8 classes but the last is always empty). As mentioned in Section 3, the individual weights w_i for each age i are built using (3), that is, by dividing 18/7 by the number of observations found in each class, except for the empty cells where the weight remains zero. Table 2 shows the results and the impact of reweighting on the derived statistics.

This example clearly illustrates the problems of aggregation bias typical of regressions estimated from economic aggregates without taking account the real distribution of the units at the micro level. Thus, it is quite clear that the units at 17 and 18 years would not have the same regression weight as those at 1 year because the risk of loss at 1 year affects almost all vehicles to be put into circulation, while very few of them will be exposed to the risk of loss of value at more advanced ages. The result is that the unweighted estimate in this example produces an over-estimation of the depreciation rate in the order of 15%.

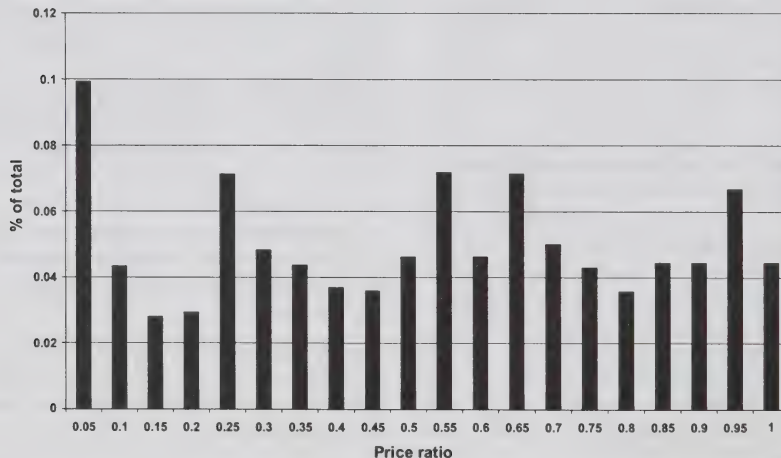


Figure 2 Weighted distribution of observations by price ratio, manufacturing plants *Ex post* weighting

Table 1 Relative prices of two models of cars based on the Kelly Blue Book and average depreciation rates before weighting

Pr ($t > S$)*		Relative price				Average depreciation rates	
Year		Excluding disposals		Including disposals		Including disposals	
		Buick	Chrysler	Buick	Chrysler	Buick	Chrysler
1	0.9988	0.8633	0.8257	0.8622	0.8246	0.1367	0.1743
2	0.9901	0.7435	0.6801	0.7361	0.6734	0.1377	0.1753
3	0.9666	0.6410	0.5608	0.6195	0.5420	0.1378	0.1754
4	0.9220	0.5523	0.4621	0.5092	0.4261	0.1379	0.1755
5	0.8526	0.4740	0.3794	0.4042	0.3234	0.1387	0.1762
6	0.7582	0.4034	0.3087	0.3058	0.2341	0.1404	0.1779
7	0.6433	0.3391	0.2482	0.2181	0.1597	0.1432	0.1805
8	0.5164	0.2790	0.1953	0.1441	0.1009	0.1475	0.1846
9	0.3892	0.2227	0.1491	0.0867	0.0580	0.1537	0.1906
10	0.2731	0.1639	0.1050	0.0448	0.0287	0.1654	0.2018
11	0.1770	0.1261	0.0772	0.0223	0.0137	0.1716	0.2077
12	0.1051	0.0892	0.0523	0.0094	0.0055	0.1824	0.2180
13	0.0567	0.0614	0.0344	0.0035	0.0019	0.1932	0.2284
14	0.0276	0.0441	0.0236	0.0012	0.0007	0.1999	0.2347
15	0.0120	0.0320	0.0164	0.0004	0.0002	0.2050	0.2396
16	0.0046	0.0190	0.0093	0.0001	0.0000	0.2194	0.2534
17	0.0016	0.0088	0.0041	0.0000	0.0000	0.2432	0.2761
18	0.0005	0.0051	0.0023	0.0000	0.0000	0.2542	0.2867
Average						0.1727	0.2087

* Survival probability based on estimates from the Micro-Economic Studies and Analysis Division of Statistics Canada.

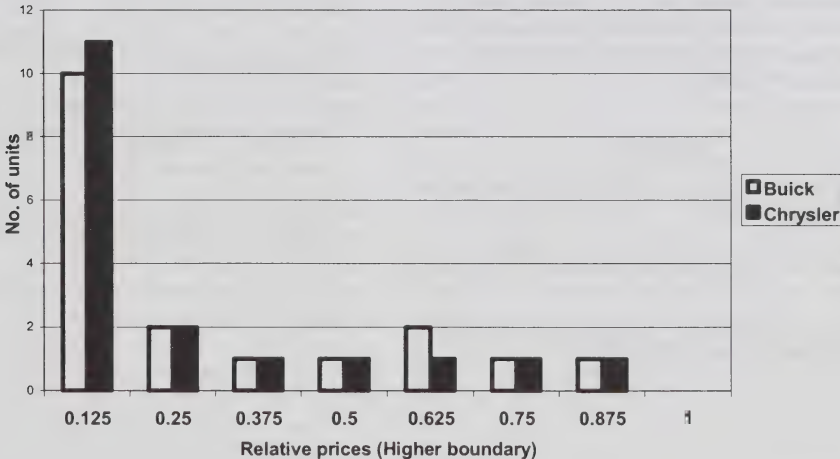


Figure 3 Distribution of cells used to estimate the average depreciation rate using data from the Kelly Blue Book before weighting (Total = 18)

Table 2 Relative prices of two models of cars based on the Kelly Blue Book and the average depreciation rate after weighting

Year	Relative prices		Average depreciation rates		<i>Ex post</i> weights	
	Including disposals		Including disposals			
	<i>Buick</i>	<i>Chrysler</i>	<i>Buick</i>	<i>Chrysler</i>	<i>Buick</i>	<i>Chrysler</i>
1	0.8622	0.8246	0.1367	0.1743	2.5714	2.5714
2	0.7361	0.6734	0.1377	0.1753	2.5714	2.5714
3	0.6195	0.5420	0.1378	0.1754	1.2857	2.5714
4	0.5092	0.4261	0.1379	0.1755	1.2857	2.5714
5	0.4042	0.3234	0.1387	0.1762	2.5714	2.5714
6	0.3058	0.2341	0.1404	0.1779	2.5714	1.2857
7	0.2181	0.1597	0.1432	0.1805	1.2857	1.2857
8	0.1441	0.1009	0.1475	0.1846	1.2857	0.2338
9	0.0867	0.0580	0.1537	0.1906	0.2571	0.2338
10	0.0448	0.0287	0.1654	0.2018	0.2571	0.2338
11	0.0223	0.0137	0.1716	0.2077	0.2571	0.2338
12	0.0094	0.0055	0.1824	0.2180	0.2571	0.2338
13	0.0035	0.0019	0.1932	0.2284	0.2571	0.2338
14	0.0012	0.0007	0.1999	0.2347	0.2571	0.2338
15	0.0004	0.0002	0.2050	0.2396	0.2571	0.2338
16	0.0001	0.0000	0.2194	0.2534	0.2571	0.2338
17	0.0000	0.0000	0.2432	0.2761	0.2571	0.2338
18	0.0000	0.0000	0.2542	0.2867	0.2571	0.2338
			<i>Weighted average</i>		0.1479	0.1836

Acknowledgements

The authors would like to express their sincere appreciation to the anonymous referee of *Survey Methodology*, whose thoughtful comments helped improve the quality of the article.

References

- Bickel, P.J., and Doksum, K.A. (1977). *Mathematical Statistics*, Holden-Day, Oakland, CA.
- Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- Davidson, R., and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, N.Y.
- Gellatly, G., Tanguay, M. and Yan, B. (2002). An alternative methodology for estimating economic depreciation: New results using a survival model. In *Productivity Growth in Canada - 2002*, Statistics Canada. #15-204-XPE.
- Greene, W.H. (1993). *Econometric Analysis*. Second edition, Prentice Hall, Englewood Cliffs, N.J.
- Hulten, C.R., and Wykoff, F.C. (1981). The measurement of economic depreciation. In *Depreciation, Inflation, and the Taxation of Income from Capital*, (Ed. C.R. Hulten). The Urban Institute Press, Washington, D.C, 81-125.
- Lancaster, T. (1985). Generalized residuals and heterogeneous duration model: With applications to the weibull model. *Journal of Econometrics*, 28, 155-69.
- Ross, S.M. (2002). *Introduction to Probability Models*, 8th Edition, Academic Press.

Person-level and household-level regression estimation in household surveys

David G. Steel and Robert G. Clark¹

Abstract

A common class of survey designs involves selecting all people within selected households. Generalized regression estimators can be calculated at either the person or household level. Implementing the estimator at the household level has the convenience of equal estimation weights for people within households. In this article the two approaches are compared theoretically and empirically for the case of simple random sampling of households and selection of all persons in each selected household. We find that the household level approach is theoretically more efficient in large samples and any empirical inefficiency in small samples is limited.

Key Words: Contextual effects; Generalized regression estimator; Intra-class correlation; Sampling variance; Model-assisted; Household surveys.

1. Introduction

Many household surveys involve selecting a sample of households and then selecting all people in the scope of the survey in the selected households. Data on one or more variables of interest are collected for the people in the sample. There may be some auxiliary variables whose population totals and sample values are known; for example these may consist of population counts by geographic and demographic classifications. The generalized regression (GREG) estimator is often used to combine auxiliary information and sample data to efficiently estimate the population totals of the variables of interest.

The GREG estimator makes use of a regression model relating the variable of interest to the auxiliary variables. The standard approach is to fit this model using data for each person in the sample (e.g., Lemaître and Dufour 1987, first paragraph). This person-level GREG estimator is equal to a weighted sum of the sample values of the variable of interest, where the weights are in general different for each person.

It is sometimes convenient to have equal weights for people within a household, for surveys which collect information on both household and person level variables of interest. The same weights can then be used for both types of variables. This ensures that relationships between household variables and person variables are reflected in estimates of total. If a household level variable is equal to the sum of person level variables (for example if household income is the sum of personal incomes), then the estimated total of the household variable will equal the estimated total of the person variable. This is not generally the case where separate weighting procedures are used for person and household variables. Similarly, if there is an inequality

relationship between a household level variable and the sum of the person level variables, this will also be reflected in the estimates of the two variables. For example, the estimated number of households using child care centres should not exceed the estimated number of children using centres.

The household-level GREG estimator achieves equal weights within households by fitting the regression model using household totals of the variable of interest and the auxiliary variables (e.g., Nieuwenbroek 1993). Weights with this property are called integrated weights.

An alternative approach would be to use different estimation methods for household-level and person-level variables, and then make an adjustment to force agreement of estimates which should be equal. This approach is sometimes called benchmarking and has mainly been used to achieve consistency between estimates from annual and sub-annual business surveys (e.g., Cholette 1984). A benchmarking approach to household and person-level variables from household surveys would require explicit identification of which person and household-level variables should have equal population totals. In this article we concentrate on integrated weighting and do not consider benchmarking approaches.

Luery (1986); Alexander (1987); Heldal (1992) and Lemaître and Dufour (1987) discussed a number of methods which give integrated weights for person-level and household-level estimates. However, none of these authors evaluated the impact on the sampling variance of calculating the generalized regression estimator at the household level rather than the person level. This is an important issue in practice because the cosmetic benefit of integrated weighting must be balanced against any effect on sampling efficiency.

1. David G. Steel and Robert G. Clark, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: David_Steel@uow.edu.au.

This article compares the design variance, which is the variance over repeated probability sampling from a fixed population, of the person-level and household-level generalized regression estimators. In Section 2, we prove that the large sample variance of the household-level estimator is less than or equal to that of the person-level estimator, by showing that the former is optimal in a large class of GREG estimators. We show that this is because the household-level estimator effectively models contextual effects whereas the person-level estimator does not. In Section 3 the two estimators are compared for a range of variables in a simulation study. Section 4 is a discussion. Three theorems are proved in an Appendix.

2. Theoretical comparison of person and household GREGs

2.1 The generalized regression estimator

In this subsection the generalized regression estimator is described for the general case of probability sampling from any population of units. Let U be a finite population of units and $s \subseteq U$ be the sample. The probabilities of selection are $\pi_i = \Pr[i \in s]$ for units $i \in U$. Let y_i be the variable of interest which is observed for units $i \in s$. Let \mathbf{z}_i be the vector of auxiliary variables for unit i , which are observed for every unit in the population. The population totals of these variables are T_y and T_z respectively.

The generalized regression estimator of T_y is based on a model relating the variable of interest to the auxiliary variables:

$$\left. \begin{aligned} E_M[y_i] &= \beta^T \mathbf{z}_i \\ \text{var}_M[y_i] &= v_i \sigma^2 \\ y_i, y_j &\text{ independent for } i \neq j \end{aligned} \right\} \quad (1)$$

where v_i are known variance parameters. Subscripts " M " refer to expectations under a model and subscripts " p " refer to design-based expectations, which are expectations over repeated probability sampling from a fixed population. For business surveys collecting continuous variables such as business income and expenses, v_i are often modelled as a function of business size. For household surveys, the variable of interest is often dichotomous, in which case v_i is usually set to 1 corresponding to a homoskedastic model.

Usually \mathbf{z}_i have the property that there exists a vector λ such that $\lambda^T \mathbf{z}_i = 1$ for all $i \in U$. For example, this is true if the regression model (1) contains an intercept parameter.

Definition 1. generalized regression estimator

The generalized regression estimator for model (1) is defined as

$$\hat{T}_r = \hat{T}_\pi + \hat{\beta}^T (T_z - \hat{T}_{z\pi}) \quad (2)$$

where

$$\begin{aligned} \hat{T}_\pi &= \sum_{i \in s} \pi_i^{-1} y_i \\ \hat{T}_{z\pi} &= \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i. \end{aligned}$$

and $\hat{\beta}$ is a solution of

$$\sum_{i \in s} c_i \pi_i^{-1} (y_i - \hat{\beta}^T \mathbf{z}_i) \mathbf{z}_i = 0$$

where c_i are regression weights. (Often c_i are set to $c_i = v_i^{-1}$.)

The coefficients $\hat{\beta}$ are calculated from a weighted least squares regression of y_i on \mathbf{z}_i for $i \in s$. The GREG estimator has low design variance if the model is approximately true but is design-consistent regardless of the truth of the model (e.g., Särndal, Swensson and Wretman 1992, chapter 6).

For large samples the design variance of \hat{T}_r is approximately equal to

$$\text{var}_p[\hat{T}_r] \approx \text{var}_p[\tilde{T}_r] \quad (3)$$

where

$$\tilde{T}_r = \hat{T}_\pi + \mathbf{B}^T (T_z - \hat{T}_{z\pi})$$

and \mathbf{B} is a solution of

$$\sum_{i \in U} c_i (y_i - \mathbf{B}^T \mathbf{z}_i) \mathbf{z}_i = 0$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). The coefficients \mathbf{B} are calculated from a weighted least squares regression of y_i on \mathbf{z}_i for $i \in U$. The sample regression coefficients $\hat{\beta}$ are design-consistent for \mathbf{B} .

2.2 Person and household level GREGs

We now consider the special case of household sampling, where the basic unit, i , is the person. Let \mathbf{x}_i be the p -vector of auxiliary variables observed for all people $i \in U$. The elements of \mathbf{x}_i may refer to characteristics of the person or of the household to which they belong. The population and sample of households will be denoted U_1 and s_1 respectively. The population of people in household $g \in U_1$ will be denoted U_g which is of size N_g . Let $y_{g1} = \sum_{i \in U_g} y_i$ and $\mathbf{x}_{g1} = \sum_{i \in U_g} \mathbf{x}_i$ be the household totals of y_i and \mathbf{x}_i . Let $\bar{\mathbf{x}}_g = \mathbf{x}_{g1} / N_g$ be the household mean of \mathbf{x}_i .

We consider the common case where households are selected by probability sampling and all people are selected from selected households, so that $s = \bigcup_{g \in s_1} U_g$. Let

$\pi_{g1} = P[g \in s_1] > 0$ be the probability of selection for household g . It follows that $\pi_i = \pi_{g1}$ for $i \in U_g$.

The person-level GREG, \hat{T}_p , is the GREG under the following model:

$$\left. \begin{aligned} E_M[y_i] &= \beta^T x_i \\ \text{var}_M[y_i] &= v_i \sigma^2 \\ y_i, y_j &\text{ independent for } i \neq j. \end{aligned} \right\} \quad (4)$$

So the person-level GREG, \hat{T}_p , is given by substituting x_i for z_i in (2). Model (4) ignores any correlations between y_i and y_j for people i and j in the same household. These correlations were 0.3 or less in most of the variables considered by Clark and Steel (2002), although higher values occurred for variables related to ethnicity, such as Indigenous self-identification. Correlations of 1 could occur for environmental variables. Tam (1995) shows that the optimal model-assisted estimator for cluster sampling is robust to mis-specification of within-cluster correlations. One way of interpreting this result is that correlations within households are not relevant to estimating population totals, because all people are selected in selected households. So within-household correlations do not help to estimate for non-sample individuals, since the sampled and non-sampled people are in distinct households.

A number of methods have been suggested for GREG-type estimation with equal weights within households. Nieuwenbroek (1993) motivated an estimator by aggregating model (4) to household level:

$$\left. \begin{aligned} E_M[y_{g1}] &= \beta^T x_{g1} \\ \text{var}_M[y_{g1}] &= v_{g1} \sigma^2 \\ y_{g1}, y_{k1} &\text{ independent for } g \neq k. \end{aligned} \right\} \quad (5)$$

where $v_{g1} = \sum_{i \in U_g} v_i$. The GREG estimator using sample data y_{g1} for $g \in s_1$ based on this model is \hat{T}_H :

$$\hat{T}_H = \hat{T}_\pi + \hat{\beta}_H^T (T_X - \hat{T}_{X\pi}) \quad (6)$$

where $\hat{\beta}_H$ is a solution of

$$\sum_{g \in s_1} \pi_{g1}^{-1} a_g (y_{g1} - \hat{\beta}_H^T x_{g1}) x_{g1} = 0. \quad (7)$$

The regression coefficient $\hat{\beta}_H$ is a household level weighted least squares regression of the sample values of y_{g1} on x_{g1} with weights $\pi_{g1}^{-1} a_g$. The values of a_g could be set to v_{g1}^{-1} . If $v_i = 1$ then $v_{g1} = N_g$ so $a_g = N_g^{-1}$. Alternatively, $a_g = 1$ could also be used.

Several other equivalent integrated weighting methods have been used. Lemaître and Dufour (1987) constructed a generalized regression estimator at person level, using \bar{x}_g instead of x_i as the auxiliary variables. Nieuwenbroek

(1993) commented that this is equivalent to (6) if $c_i = a_g N_g$ for $i \in U_g$. Alexander (1987) developed closely related weighting methods using a minimum distance criterion.

Both the person and household level GREG can be written in weighted form $\sum_{i \in s} w_i Y_i$. The weights for both estimators can be written as $w_i = \pi_i^{-1} g_i$ where

$$g_i = 1 + (T_X - \hat{T}_{X\pi})^T \left(\sum_{i \in s} c_i \pi_i^{-1} x_i x_i^T \right)^{-1} c_i x_i$$

for \hat{T}_p and

$$g_i = 1 + (T_X - \hat{T}_{X\pi})^T \left(\sum_{g \in s_1} a_g \pi_{g1}^{-1} x_{g1} x_{g1}^T \right)^{-1} a_g \pi_{g1}^{-1} x_{g1}$$

for \hat{T}_H , where person i belongs to household g . (Superscript “ $-$ ” stands for generalized inverse of a matrix).

2.3 Theoretical results

In this section, we show that \hat{T}_H has the lowest possible large sample variance in a class of estimators which also includes \hat{T}_p , for the sample design where households are selected by simple random sampling without replacement. We will then explain this result by showing that \hat{T}_H is equivalent to a regression estimator calculated using person level data, where the model includes contextual effects.

For large samples, \hat{T}_p and \hat{T}_H can be approximated by

$$\tilde{T}_p = \hat{T}_\pi + B_p^T (T_X - \hat{T}_{X\pi});$$

and

$$\tilde{T}_H = \hat{T}_\pi + B_H^T (T_X - \hat{T}_{X\pi})$$

respectively, where B_p and B_H are solutions of

$$\left. \begin{aligned} \sum_{i \in U} c_i (y_i - B_p^T x_i) x_i &= 0 \\ \sum_{g \in U_1} a_g (y_{g1} - B_H^T x_{g1}) x_{g1} &= 0 \end{aligned} \right\} \quad (8)$$

(Särndal *et al.* 1992, Result 6.6.1, page 235). Theorem 1 states the minimum variance estimator in a class including \tilde{T}_p and \tilde{T}_H .

Theorem 1. Optimal estimator for simple cluster sampling

Suppose that m households are selected by simple random sampling without replacement from a population of M households, and all people are selected from selected households. Consider the estimator of T given by

$$\tilde{T} = \hat{T}_\pi + h^T (T_X - \hat{T}_{X\pi})$$

where h is a constant p -vector. It is assumed that there exists a vector λ such that $\lambda^T x_i = 1$ for all $i \in U$. The

variance of this estimator is minimised by h^* which are solutions of

$$\sum_{g \in s_1} (y_{g1} - h^T x_{g1}) x_{g1} = 0.$$

Hence \hat{T}_H with $a_g = 1$ for all g is the optimal choice of \hat{T} .

Theorem 1 has the perhaps surprising implication that \hat{T}_H (with $a_g = 1$ for all g) has lower variance than \hat{T}_P for large samples. This is in spite of the fact that \hat{T}_H discards some of the information in the sample, because it uses the household sums of x_i and y_i . The Theorem suggests that \hat{T}_H is the appropriate GREG estimator for the cluster sampling design assumed here, and that the information discarded by summing to household level is not relevant when this design is used. To explain why \hat{T}_H can perform better than \hat{T}_P , we will make use of a "linear contextual model" which is a more general model for $E_M[Y_i]$ than (4). The model is:

$$\left. \begin{aligned} E_M[Y_i] &= \gamma_1^T \bar{x}_g + \gamma_2^T x_i \quad (i \in U_g) \\ \text{var}_M[Y_i] &= \sigma^2 \\ y_i, y_j &\text{ independent for } i \neq j. \end{aligned} \right\} \quad (9)$$

Both \bar{x}_g and x_i are used as explanatory variables for y_i because the household mean of the person level auxiliary variables may capture some of the effect of household context (Lazarfeld and Menzel 1961). For example, if the elements of x_i are indicator variables summarising the age and sex of person i then \bar{x}_g are the proportions of people in the household falling into different age and sex categories. If the population of interest includes both adults and children, then \bar{x}_g includes the proportion of children in the household, which could be relevant to the labour force participation of adults in the household.

Theorem 2 shows that the improvement in the variance from using \hat{T}_H with $a_g = 1$ rather than using \hat{T}_P can be explained by the linear contextual model.

Theorem 2. Explaining the difference in the asymptotic variances

Suppose that households are selected by simple random sampling without replacement and all people are selected from selected households. Let $r_i = y_i - B_P^T x_i$, and let B_C be the result of regressing r_i on \bar{x}_g over $i \in U$ using weighted least squares regression weighted by N_g . Then

$$\begin{aligned} \text{var}_p[\tilde{T}_P] - \text{var}_p[\tilde{T}_H] &= \\ \frac{M^2}{m} (1 - \frac{m}{M}) (M - 1)^{-1} B_C^T &\left(\sum_{g \in U_1} x_{g1} x_{g1}^T \right) B_C \end{aligned}$$

where \tilde{T}_H is calculated using $a_g = 1$ for all g .

The result shows that the reduction in variance from using \hat{T}_H (with $a_g = 1$) rather than \hat{T}_P is a quadratic form in B_C . Hence the extent of the improvement depends on the extent to which \bar{x}_g helps to predict y_i after x_i has already been controlled for, *i.e.*, the extent to which a linear contextual effect helps to predict r_i over $i \in U$, using a weighted least squares regression weighted by N_g .

The proofs of Theorems 1 and 2 are very much dependent on the assumption of cluster sampling. The results would not be expected to apply if there was subsampling within households.

Theorems 1 and 2 only apply with $a_g = 1$ in the weighted least squares regression for \hat{T}_H . Other choices of a_g are often used, for example it would often be reasonable to assume that $v_{g1} = N_g$ in model (5), in which case it would be sensible to use $a_g = N_g^{-1}$. Theorem 3 shows that \hat{T}_H is equivalent to a person-level GREG estimator fitted under the linear contextual model for other choices of a_g .

Theorem 3. The linear contextual GREG

For sample designs where all people are selected from selected households and $\pi_{g1} > 0$ for all $g \in U$, \hat{T}_H with a given choice of a_g is the generalized regression estimator for model (9) where $c_i = a_g N_g$ for $i \in U_g$.

Theorem 3 means that \hat{T}_H is the GREG under a more general model than \hat{T}_P . Nieuwenbroek (1993) showed that \hat{T}_H is equal to a person-level GREG derived from regressing y_i on \bar{x}_g . Theorem 3 states it is also equal to the person-level GREG from regressing y_i on both x_i and \bar{x}_g , thereby automatically incorporating any household contextual effects. As a result, \hat{T}_H would be expected to have lower variance than \hat{T}_P for large samples. (In the case of $a_g = 1$, Theorem 1 stated that this is always the case). For small samples, however, a more general model may be counter-productive. Silva and Skinner (1997) showed for single-stage sampling that adding parameters to the model can increase the variance of the GREG estimator, although this effect is negligible for large samples. It is possible that the contextual effects have little or no predictive power for some variables. In this case, it would be expected that \hat{T}_H would perform slightly worse than \hat{T}_P for small samples, and about the same for large samples.

The contextual model, (9), includes all of the elements of x_i and all of the elements of \bar{x}_g . An alternative would be to use only those elements of either x_i and \bar{x}_g which are significant, or which give improvements in the estimated variance of a GREG estimator. A GREG estimator based on

this type of model would probably have lower variance than the estimators considered in this paper, but would not give integrated weights unless the same elements of x_i and \bar{x}_g were used.

3. Empirical study

3.1 Methodology

A simulation study was undertaken to compare the person and household GREGs, \hat{T}_p and \hat{T}_H , for a range of survey variables. We used two populations, consisting of 187,178 households randomly selected from the 2001 Australian Population Census and 210,132 households from the 1995 Australian National Health Survey. All adults and children in the households were included. The average household size was approximately 2.5.

We selected cluster samples from these populations, where households were selected by simple random sampling without replacement and all people from selected households were selected. We simulated samples of size $m = 500, 1,000, 2,000, 5,000$ and $10,000$ households. In each case, 5,000 samples were selected. The auxiliary variables x_i consisted of indicator variables of sex by agegroup (12 categories). (This choice of x_i means that the GREG estimation is equivalent to post-stratification.) The person-level GREG with $c_i = 1(\hat{T}_p)$, the household-level GREG with $a_g = N_g^{-1}(\hat{T}_{H1})$, and the household-level GREG with $a_g = 1(\hat{T}_{H2})$ were all calculated. We also included the Hájek estimator

$$\hat{T}_1 = N \left(\sum_{i \in S} \pi_i^{-1} y_i \right) / \left(\sum_{i \in S} \pi_i^{-1} \right)$$

which equals $N/n \sum_{g \in S} \sum_{i \in U_g} y_i$ for cluster sampling with simple random sampling of households, where n is the realized sample size of people.

The variables include labour force, health and other topics. All of the variables are dichotomous except for income (annual income in Australian dollars, based on range data reported from the Census). "Employment(F)" is the indicator variable which is 1 if a person is employed and female, and 0 otherwise. The first six variables are from the Census population and the remaining five variables are from the health population.

3.2 Results

Table 1 shows the relative root mean squared errors (RRMSEs) of \hat{T}_1 , \hat{T}_p , \hat{T}_{H1} and \hat{T}_{H2} , for a sample size of 1,000 households. The RRMSEs are expressed as a percentage of the true population total. The biases have not been tabulated because they were a negligible component of the MSE in all cases. The percentage improvements in MSE

of \hat{T}_{H1} and \hat{T}_{H2} relative to \hat{T}_p are also shown. The figures in brackets are the simulation standard errors of these percentage improvements.

For this sample size, \hat{T}_{H1} and \hat{T}_{H2} performed slightly worse than \hat{T}_p for the health variables and slightly better for most other variables. The greatest gain was in estimating the number of sole parents; this variance was reduced by 10.8% and 16.3% by using the household-level GREGs. For all other variables, either the improvement was small or the household GREG was slightly worse than the person-level GREG. The inefficiency from using a household-level GREG rather than \hat{T}_p was never more than 2.2%.

Table 2 shows the percentage improvement in MSE from using \hat{T}_{H1} rather than \hat{T}_p for different sample sizes. The simulation standard errors for each figure are shown in brackets. Table 3 shows the percentage improvements from using \hat{T}_{H2} rather than \hat{T}_p . The asymptotic percentage improvements ($m = \infty$) are also shown, based on the large sample approximation to the variance of a GREG. For both household-level GREGs, the percentage improvements are generally increasing as the sample size increases. For $m = 500$, the household GREGs are generally worse than the person GREGs, although never more than 5% worse. For $m = 10,000$, an improvement is recorded for over half of the variables. The greatest improvements were for estimates of the number of sole parents (11.5%) and employed women (4.2%); all other improvements were small. \hat{T}_{H1} and \hat{T}_{H2} never had variances more than 0.2% higher than \hat{T}_p for $m = 10,000$. Generally \hat{T}_{H2} performs better than \hat{T}_{H1} for larger sample sizes, as would be expected from Theorem 1, but the reverse is true for small sample sizes.

In practice estimates of subpopulation totals are often of as much interest as population totals. Table 4 shows the performance of the various estimators for age-sex domains (12 age categories) and region domains, for the sample size of 1,000 households. There were 49 regions in the census dataset. The health dataset did not contain a similar region variable, instead the socioeconomic quintile of the collection district (a geographical unit consisting of approximately 200 contiguous households) was used as the domain. The domain estimators were produced by calculating weights from each estimator and taking the weighted sum over the sample in the domain. This is equivalent to the domain ratio estimator described in Case 1, Section 2.1 of Hidiroglou and Patak (2004). We have used this method because it is the most commonly used in practice, as it enables all domains and population totals to be estimated with a single set of weights, although more efficient domain estimators exist (Hidiroglou and Patak 2004, cases 2-6).

In each case, the median RRMSE over the domains is shown. The table shows that there is not much difference

between the three GREG estimators. For age-sex domains, the household GREGs did slightly better than the person GREG for census variables and slightly worse for health variables. For region estimates, the household GREGs were slightly worse in all cases. Table 5 shows that the

households GREGs performed very similarly to \hat{T}_P for a sample size of 10,000 households. It is worth noting that Theorem 1 and 2 do not apply to the domain estimators we have used.

Table 1 Relative RMSEs for sample size of 1,000 households

Variable	RRMSE%				% improvement in MSE	
	\hat{T}_1	\hat{T}_P	\hat{T}_{H1}	\hat{T}_{H2}	\hat{T}_{H1}	\hat{T}_{H2}
employed	2.62	2.09	2.09	2.10	0.20 (0.26)	-0.28 (0.27)
employed F	3.78	3.05	3.01	3.02	2.63 (0.33)	2.09 (0.33)
income	2.56	2.20	2.19	2.19	1.04 (0.25)	0.75 (0.24)
low income	5.04	4.87	4.89	4.90	-0.62 (0.20)	-1.12 (0.22)
hrs worked	3.08	2.54	2.53	2.53	0.94 (0.28)	0.70 (0.28)
sole parent	12.50	12.73	12.02	11.65	10.84 (0.62)	16.31 (0.49)
arthritis	5.52	4.50	4.53	4.53	-1.38 (0.17)	-1.57 (0.18)
smoker	4.73	4.57	4.60	4.61	-1.64 (0.18)	-1.81 (0.20)
high BPR	6.80	5.30	5.35	5.36	-1.70 (0.17)	-2.06 (0.18)
fair/poor hlth	9.79	9.42	9.47	9.47	-1.16 (0.16)	-1.07 (0.18)
alcohol	4.81	4.66	4.70	4.71	-1.77 (0.16)	-2.15 (0.18)

Table 2 Improvement in MSE of household GREG \hat{T}_{H1} compared to \hat{T}_P

Variable	% improvement in MSE					
	$m = 500$	1,000	2,000	5,000	10,000	∞
employed	-0.65 (0.31)	0.20 (0.26)	1.02 (0.24)	0.90 (0.21)	2.17 (0.21)	1.85
employed F	1.22 (0.37)	2.63 (0.33)	2.59 (0.33)	3.53 (0.31)	4.24 (0.31)	4.13
income	-1.53 (0.31)	1.04 (0.25)	0.48 (0.24)	0.61 (0.19)	1.43 (0.19)	1.07
low income	-2.45 (0.27)	-0.62 (0.20)	0.02 (0.18)	0.18 (0.15)	0.00 (0.00)	0.65
hrs worked	-0.26 (0.34)	0.94 (0.28)	1.72 (0.27)	1.61 (0.24)	2.64 (0.24)	2.12
sole parent	7.81 (0.69)	10.84 (0.62)	10.74 (0.61)	10.23 (0.57)	11.50 (0.58)	11.21
arthritis	-3.01 (0.24)	-1.38 (0.17)	-0.34 (0.12)	-0.08 (0.09)	-0.13 (0.07)	0.08
smoker	-3.91 (0.25)	-1.64 (0.18)	-1.02 (0.12)	-0.26 (0.08)	-0.06 (0.07)	0.16
high BPR	-2.93 (0.24)	-1.70 (0.17)	-0.86 (0.12)	-0.31 (0.08)	-0.04 (0.06)	0.08
fair/poor hlth	-3.67 (0.25)	-1.16 (0.16)	-0.71 (0.12)	-0.05 (0.08)	0.03 (0.06)	0.10
alcohol	-4.22 (0.23)	-1.77 (0.16)	-0.77 (0.12)	-0.31 (0.08)	-0.21 (0.07)	0.14

Table 3 Improvement in MSE of household GREG \hat{T}_{H2} compared to \hat{T}_P

Variable	% improvement in MSE					
	$m = 500$	1,000	2,000	5,000	10,000	∞
employed	-1.85 (0.35)	-0.28 (0.27)	1.25 (0.25)	1.05 (0.21)	2.22 (0.21)	1.98
employed F	0.28 (0.39)	2.09 (0.33)	2.71 (0.33)	3.55 (0.29)	4.50 (0.30)	4.31
income	-2.64 (0.31)	0.75 (0.24)	0.71 (0.22)	0.90 (0.17)	1.30 (0.16)	1.37
low income	-3.15 (0.30)	-1.12 (0.22)	-0.15 (0.18)	0.06 (0.15)	0.00 (0.00)	0.94
hrs worked	-1.51 (0.35)	0.70 (0.28)	1.98 (0.25)	1.79 (0.21)	2.57 (0.22)	2.26
sole parent	14.70 (0.53)	16.31 (0.49)	16.39 (0.47)	15.41 (0.44)	16.44 (0.44)	16.35
arthritis	-3.31 (0.26)	-1.57 (0.18)	-0.05 (0.13)	-0.12 (0.09)	-0.10 (0.07)	0.16
smoker	-3.82 (0.28)	-1.81 (0.20)	-0.69 (0.14)	0.21 (0.11)	0.28 (0.10)	0.57
high BPR	-3.20 (0.26)	-2.06 (0.18)	-1.12 (0.13)	-0.40 (0.09)	-0.05 (0.07)	0.12
fair/poor hlth	-4.02 (0.28)	-1.07 (0.18)	-0.57 (0.13)	-0.09 (0.09)	0.00 (0.07)	0.15
alcohol	-5.00 (0.26)	-2.15 (0.18)	-0.82 (0.13)	-0.49 (0.09)	-0.29 (0.08)	0.18

Table 4 Median relative RMSEs for domain estimators for sample size $m = 1,000$

Variable	Age-Sex Domains				Region Domains			
	\hat{T}_1	\hat{T}_P	\hat{T}_{H1}	\hat{T}_{H2}	\hat{T}_1	\hat{T}_P	\hat{T}_{H1}	\hat{T}_{H2}
employed	12.74	7.92	7.93	7.90	29.89	29.92	30.20	30.34
employed F	13.12	8.32	8.36	8.34	34.64	34.65	35.03	35.16
income	13.25	8.43	8.49	8.47	28.04	28.12	28.43	28.51
low income	21.17	18.77	18.96	18.94	42.71	42.85	43.24	43.33
hrs worked	14.56	10.69	10.76	10.72	31.24	31.23	31.52	31.63
sole parent	96.20	96.33	97.64	96.69	92.99	93.30	94.37	93.50
arthritis	24.94	20.94	21.12	21.11	13.31	12.94	13.02	13.04
smoker	32.10	29.25	29.39	29.37	12.32	12.27	12.35	12.38
high BPR	27.01	23.80	23.97	23.95	15.83	15.31	15.44	15.45
fair/poor hlth	39.64	37.73	38.05	38.08	22.38	22.30	22.51	22.55
alcohol	25.58	21.42	21.53	21.58	12.73	12.70	12.80	12.82

Table 5 Median relative RMSEs for domain estimators for sample size $m = 10,000$

Variable	Age-Sex Domains				Region Domains			
	\hat{T}_1	\hat{T}_P	\hat{T}_{H1}	\hat{T}_{H2}	\hat{T}_1	\hat{T}_P	\hat{T}_{H1}	\hat{T}_{H2}
employed	3.77	2.35	2.32	2.31	8.85	8.85	8.87	8.88
employed F	3.86	2.43	2.43	2.42	10.30	10.26	10.25	10.25
income	3.91	2.53	2.51	2.51	8.24	8.23	8.23	8.24
low income	6.31	5.63	5.62	5.61	12.67	12.68	12.69	12.69
hrs worked	4.29	3.15	3.15	3.12	9.26	9.25	9.27	9.27
sole parent	28.40	28.26	28.29	28.23	27.11	27.14	27.16	27.11
arthritis	7.40	6.26	6.27	6.27	3.98	3.85	3.85	3.85
smoker	9.53	8.58	8.58	8.57	3.69	3.67	3.68	3.67
high BPR	8.07	7.02	7.01	7.01	4.66	4.48	4.49	4.49
fair/poor hlth	11.69	11.02	11.02	11.01	6.75	6.69	6.69	6.69
alcohol	7.74	6.43	6.43	6.43	3.87	3.85	3.85	3.85

4. Discussion

The standard person-level GREG estimator produces unequal weights within households. Household-level GREG estimators can be used to give integrated household and person weights, which is beneficial for surveys collecting information on both household-level and person-level variables. This article demonstrated that there is little or no loss associated with the practical benefit of integrated weighting arising from using a household-level GREG estimator. For large samples, the household-level GREG has lower design variance than the person-level GREG. For smaller samples there is at most a small increase in variance for some variables from using the household GREG, because this estimator is equivalent to using a regression model containing more parameters. Therefore, if integrated weights would improve the coherence of a household survey's outputs, the household-level GREG can be adopted with little or no detriment to the variance and bias of estimators.

Acknowledgements

This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics. The views expressed here do not necessarily reflect the views of either organisation. The authors thank Julian England, Frank Yu and Ray Chambers for their thoughtful comments.

Appendix

Proof of theorems

Proof of theorem 1

Let $\bar{Y}_1 = T_Y/M$ and $\bar{X}_1 = T_X/M$ be the population means of y_{g1} and x_{g1} respectively. The variance of \tilde{T} is

$$\begin{aligned} \text{var}_p[\tilde{T}] &= \text{var}[\hat{T}_\pi + h^T(T_X - \hat{T}_{X\pi})] \\ &= \text{var}\left[\frac{M}{m} \sum_{g \in s_1} (y_{g1} - h^T x_{g1})\right] \\ &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) S_r^2 \end{aligned}$$

where $S_r^2 = (M-1)^{-1} \sum_{g \in U_i} \{y_{g1} - \mathbf{h}^T \mathbf{x}_{g1} - (\bar{Y}_1 - \mathbf{h}^T \bar{X}_1)\}^2$. To minimise with respect to \mathbf{h} , we set the derivative of S_r^2 to zero:

$$\begin{aligned} 0 &= (M-1)^{-1} \sum_{g \in U_i} \{y_{g1} - \mathbf{h}^T \mathbf{x}_{g1} - (\bar{Y}_1 - \mathbf{h}^T \bar{X}_1)\} (\mathbf{x}_{g1} - \bar{X}_1) \\ 0 &= \sum_{g \in U_i} \{y_{g1} - \mathbf{h}^T \mathbf{x}_{g1} - (\bar{Y}_1 - \mathbf{h}^T \bar{X}_1)\} \mathbf{x}_{g1} \\ &\quad - \sum_{g \in U_i} \{y_{g1} - \bar{Y}_1 - \mathbf{h}^T (\mathbf{x}_{g1} - \bar{X}_1)\} \bar{X}_1 \\ 0 &= \sum_{g \in U_i} \{y_{g1} - \mathbf{h}^T \mathbf{x}_{g1} - (\bar{Y}_1 - \mathbf{h}^T \bar{X}_1)\} \mathbf{x}_{g1} \\ 0 &= \sum_{g \in U_i} (y_{g1} - \mathbf{h}^T \mathbf{x}_{g1}) \mathbf{x}_{g1} - (\bar{Y}_1 - \mathbf{h}^T \bar{X}_1) T_X. \quad (10) \end{aligned}$$

We now show that (10) is satisfied by \mathbf{h}^* . By assumption, \mathbf{h}^* satisfies

$$\mathbf{0} = \sum_{g \in U_i} (y_{g1} - \mathbf{x}_{g1}^T \mathbf{h}^*) \mathbf{x}_{g1}. \quad (11)$$

Hence the first sum in the right hand side of (10) is equal to zero for $\mathbf{h} = \mathbf{h}^*$. Premultiplying both sides of (11) by λ^T gives

$$\begin{aligned} 0 &= \sum_{g \in U_i} (y_{g1} - \mathbf{x}_{g1}^T \mathbf{h}^*) \lambda^T \mathbf{x}_{g1} \\ 0 &= \sum_{g \in U_i} (y_{g1} - \mathbf{x}_{g1}^T \mathbf{h}^*) \\ 0 &= T_Y - T_X^T \mathbf{h}^*. \end{aligned}$$

Dividing by M gives $\bar{Y}_1 - \bar{X}_1^T \mathbf{h}^* = 0$. Hence the rest of the right hand side of (10) is equal to zero. So \mathbf{h}^* satisfies (10).

Proof of theorem 2

Let “ \cdot ” denote a generalized inverse of a matrix. Then B_C is equal to

$$\begin{aligned} B_C &= \left\{ \sum_{g \in U_i} \sum_{i \in U_g} N_g \bar{x}_g \bar{x}_g^T \right\}^{-1} \sum_{g \in U_i} \sum_{i \in U_g} N_g \bar{x}_g r_i \\ &= \left\{ \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \right\}^{-1} \sum_{g \in U_i} \mathbf{x}_{g1} r_{g1}. \quad (12) \end{aligned}$$

Now, $r_i = y_i - \mathbf{B}_p^T \mathbf{x}_i$ so $r_{g1} = y_{g1} - \mathbf{B}_p^T \mathbf{x}_{g1}$. Hence (12) becomes

$$\begin{aligned} B_C &= \left\{ \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \right\}^{-1} \sum_{g \in U_i} \mathbf{x}_{g1} (y_{g1} - \mathbf{B}_p^T \mathbf{x}_{g1}) \\ &= \left\{ \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \right\}^{-1} \sum_{g \in U_i} \mathbf{x}_{g1} y_{g1} \\ &\quad - \left\{ \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \right\}^{-1} \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \mathbf{B}_p \\ &= B_H - B_p \quad (13) \end{aligned}$$

since $B_H = \left\{ \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \right\}^{-1} \sum_{g \in U_i} \sum_{i \in U_g} \mathbf{x}_{g1} y_{g1}$. The difference in the variances is given by

$$\begin{aligned} \text{var}_p[\tilde{T}_p] - \text{var}_p[\tilde{T}_H] &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) (M-1)^{-1} \\ &\quad \left\{ \sum_{g \in U_i} (y_{g1} - \mathbf{B}_p^T \mathbf{x}_{g1})^2 - \sum_{g \in U_i} (y_{g1} - \mathbf{B}_H^T \mathbf{x}_{g1})^2 \right\} \end{aligned}$$

which becomes

$$\begin{aligned} &\left\{ \text{var}_p[\tilde{T}_p] - \text{var}_p[\tilde{T}_H] \right\} \left/ \left\{ \frac{M^2}{m} \left(1 - \frac{m}{M}\right) (M-1)^{-1} \right\} \right. \\ &= \sum_{g \in U_i} r_{g1}^2 - \sum_{g \in U_i} (r_{g1} + \mathbf{B}_p^T \mathbf{x}_{g1} - \mathbf{B}_H^T \mathbf{x}_{g1})^2 \\ &= \sum_{g \in U_i} r_{g1}^2 - \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1})^2 \\ &= \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1} + \mathbf{B}_C^T \mathbf{x}_{g1})^2 - \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1})^2 \\ &= \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1})^2 + \sum_{g \in U_i} (\mathbf{B}_C^T \mathbf{x}_{g1})^2 \\ &\quad + 2 \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1}) \mathbf{x}_{g1}^T \mathbf{B}_C^T \\ &\quad - \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1})^2 \\ &= \sum_{g \in U_i} \mathbf{B}_C^T \mathbf{x}_{g1} \mathbf{x}_{g1}^T \mathbf{B}_C + 2 \sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1}) \mathbf{x}_{g1}^T \mathbf{B}_C. \quad (14) \end{aligned}$$

Now, B_C is an ordinary least squares regression of r_{g1} on \mathbf{x}_{g1} so

$$\sum_{g \in U_i} (r_{g1} - \mathbf{B}_C^T \mathbf{x}_{g1}) \mathbf{x}_{g1} = \mathbf{0}.$$

Hence (14) becomes

$$\begin{aligned} \text{var}_p[\tilde{T}_p] - \text{var}_p[\tilde{T}_H] &= \\ &\frac{M^2}{m} \left(1 - \frac{m}{M}\right) (M-1)^{-1} \mathbf{B}_C^T \sum_{g \in U_i} \mathbf{x}_{g1} \mathbf{x}_{g1}^T \mathbf{B}_C. \end{aligned}$$

Proof of Theorem 3

The GREG estimator is invariant under linear invertible transformations of the auxiliary variables. Hence model (9) can be re-parameterised to give

$$E_M[y_i] = \phi_1^T \bar{x}_g + \phi_2^T (x_i - \bar{x}_g) \quad (15)$$

or equivalently

$$E_M[y_i] = \phi^T z_i$$

where

$$z_i = \begin{pmatrix} \bar{x}_g \\ x_i - \bar{x}_g \end{pmatrix}$$

and

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}.$$

The parameters in model (15) are related to those in model (9) by $\phi_1 = \gamma_1 + \gamma_2$ and $\phi_2 = \gamma_2$.

From Definition 1, noting that

$$s = \bigcup_{g \in s_1} U_g$$

for the assumed design, the generalized regression estimator under model (15) is

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{i \in U} \hat{\phi}^T z_i - \sum_{i \in s} \pi_i^{-1} \hat{\phi}^T z_i \\ &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \{ \hat{\phi}_1^T \bar{x}_g + \hat{\phi}_2^T (x_i - \bar{x}_g) \} \\ &\quad - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \{ \hat{\phi}_1^T \bar{x}_g + \hat{\phi}_2^T (x_i - \bar{x}_g) \}. \end{aligned} \quad (16)$$

However, $\sum_{i \in U_g} (x_i - \bar{x}_g) = 0$ for each g . Hence (16) becomes

$$\begin{aligned} \hat{T} &= \hat{T}_\pi + \sum_{g \in U_1} \sum_{i \in U_g} \hat{\phi}_1^T \bar{x}_g - \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} \hat{\phi}_1^T \bar{x}_g \\ &= \hat{T}_\pi + \hat{\phi}_1^T \sum_{g \in U_1} \sum_{i \in U_g} \bar{x}_g - \hat{\phi}_1^T \sum_{g \in s_1} \pi_g^{-1} \sum_{i \in U_g} \bar{x}_g \\ &= \hat{T}_\pi + \hat{\phi}_1^T \sum_{g \in U_1} \bar{x}_{g1} - \hat{\phi}_1^T \sum_{g \in s_1} \pi_{g1}^{-1} \bar{x}_{g1} \\ &= \hat{T}_\pi + \hat{\phi}_1^T (T_X - \hat{T}_{X\pi}). \end{aligned} \quad (17)$$

Notice that (17) does not include the estimator of ϕ_2 . The least squares estimators

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix}$$

are the solution of:

$$\sum_{i \in s} \pi_i^{-1} c_i (y_i - \hat{\phi}^T z_i) z_i = 0$$

which is equivalent to:

$$\sum_{i \in s} \pi_i^{-1} c_i \{ y_i - \hat{\phi}_1^T \bar{x}_g - \hat{\phi}_2^T (x_i - \bar{x}_g) \} \begin{pmatrix} \bar{x}_g \\ x_i - \bar{x}_g \end{pmatrix} = 0.$$

By assumption, $c_i = a_g N_g$ so the first p elements of this equation are:

$$0 = \sum_{g \in s_1} \sum_{i \in U_g} \pi_i^{-1} a_g N_g \bar{x}_g \{ y_i - \hat{\phi}_1^T \bar{x}_g - \hat{\phi}_2^T (x_i - \bar{x}_g) \}$$

$$0 = \sum_{g \in s_1} \pi_{g1}^{-1} a_g N_g \bar{x}_g \sum_{i \in U_g} \{ y_i - \hat{\phi}_1^T \bar{x}_g - \hat{\phi}_2^T (x_i - \bar{x}_g) \}$$

$$0 = \sum_{g \in s_1} \pi_{g1}^{-1} a_g x_{g1} \{ y_{g1} - \hat{\phi}_1^T x_{g1} - \hat{\phi}_2^T (x_{g1} - x_{g1}) \}$$

$$0 = \sum_{g \in s_1} \pi_{g1}^{-1} a_g x_{g1} (y_{g1} - \hat{\phi}_1^T x_{g1}).$$

Hence $\hat{\phi}_1$ is a solution to (7). So the GREG estimator for model (9) is equal to \hat{T}_H provided that $c_i = a_g N_g$.

References

- Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- Cholette, P. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10, 35-49.
- Clark, R.G., and Steel, D.G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, 70 (2), 289-314.
- Heldal, J. (1992). A method for calibration of weights in sample surveys. In *Workshop on uses of auxiliary information in surveys*. University of Orebro, Sweden.
- Hidiroglou, M., and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 67-78.
- Lazarfeld, P.F., and Menzel, H. (1961). On the relation between individual and collective properties. In *Complex Organizations: A Sociological Reader*. Holt, Reinhart and Winston. 422-440.
- Lemaitre, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- Luery, D.M. (1986). Weighting sample survey data under linear constraints on the weights. In *Proceedings of the Social Statistics Section*, American Statistical Association, (Alexandria, VA), 325-330.
- Nieuwenbroek, N. (1993). *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Netherlands Central Bureau of Statistics.

Särndal, C., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Silva, P.L.N., and Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.

Mean - Adjusted bootstrap for two - Phase sampling

Hiroshi Saigo ¹

Abstract

Two-phase sampling is a useful design when the auxiliary variables are unavailable in advance. Variance estimation under this design, however, is complicated particularly when sampling fractions are high. This article addresses a simple bootstrap method for two-phase simple random sampling without replacement at each phase with high sampling fractions. It works for the estimation of distribution functions and quantiles since no rescaling is performed. The method can be extended to stratified two-phase sampling by independently repeating the proposed procedure in different strata. Variance estimation of some conventional estimators, such as the ratio and regression estimators, is studied for illustration. A simulation study is conducted to compare the proposed method with existing variance estimators for estimating distribution functions and quantiles.

Key Words: Double Sampling; Resampling; Variance estimation.

1. Introduction

Two-phase sampling or double sampling is a powerful tool for efficient estimation in surveys. Usually, a large-scale first phase sample is taken where auxiliary variables, correlated with the characteristics of interest and relatively easily obtained, are observed. Then, a small-scale sub-sample is chosen from the first phase sample to measure the characteristics of interest that are harder to obtain. At the estimation stage, the auxiliary variables at the first phase are employed to obtain an efficient estimator.

A closed-form sample variance formula for an estimator can be complicated or even unavailable under two-phase sampling. Consequently, resampling methods, such as the jackknife and bootstrap, are appealing for two-phase sampling. Rao and Sitter (1995) and Sitter (1997) studied the delete-1 jackknife approach to the ratio and regression estimators under two-phase sampling and found the method provides design-consistent variance estimation with desirable conditional properties given the auxiliary variables.

A weakness of the delete-1 jackknife is that it cannot handle quantile estimation. Moreover, it is not trivial how one can incorporate the finite population correction into the jackknife variance estimation under two-phase sampling (see Lee and Kim 2002 and Berger and Rao 2006). The bootstrap, on the other hand, eliminates these problems if properly formulated.

Several bootstrap methods for two-phase sampling have been proposed and studied. Schreuder, Li and Scott (1987), Biemer and Atkinson (1993) and Sitter (1997) considered similar bootstrap methods which provide consistent variance estimation when sampling fractions are negligible. Rao and Sitter (1997) proposed a rescaling bootstrap for high sampling fractions.

A disadvantage of the rescaling approach is that it cannot handle the estimation of distribution functions and quantiles. In this paper, we propose a mean-adjusted bootstrap for two-phase sampling that accommodates the estimation of distribution functions and quantiles. The method is simple and includes the existing ones for negligible sampling fractions as a special case. Recently, Kim, Navarro, and Fuller (2006) studied replication variance estimation without rescaling for two-phase sampling in a more generalized framework than that of this paper. Our method, however, is different in that it internally incorporates the finite population correction.

This paper is organized as follows. Section 2 presents the mean-adjusted bootstrap for two-phase sampling. Section 3 illustrates how the proposed method works for some conventional estimators. A simulation for estimating distribution functions and quantiles is conducted in Section 4. Section 5 discusses further applications of the mean-adjusted bootstrap. Concluding remarks are given in Section 6.

2. Mean - Adjusted bootstrap

For notational simplicity, we assume there is only one stratum. To extend our method to stratified sampling, repeat the same procedure independently in different strata to obtain a bootstrap sample (see Rao and Sitter 1997, pages 759-762).

Let P be the set of unit labels in a population of size N . Suppose a simple random sample without replacement (SRSWOR) of size n_{A+B} from P is taken and denote the sampled labels by $A+B$. The auxiliary variable (vector) x_i is observed for $i \in A+B$. Then take a second phase SRSWOR of size $n_A < n_{A+B}$ from $A+B$ and denote the sampled labels by A . The characteristic (vector) y_i is

measured for $i \in A$. Let $B = (A + B) - A$, $n_B = n_{A+B} - n_A$, $\mathbf{y}_A = \{y_i : i \in A\}$, $\mathbf{x}_A = \{x_i : i \in A\}$, and $\mathbf{x}_B = \{x_j : j \in B\}$. An approximately design-unbiased estimator of parameter θ is assumed to be written as $\hat{\theta} = t(\mathbf{y}_A, \mathbf{x}_A, \mathbf{x}_B)$.

Under the proposed method, a bootstrap sample is constructed as follows.

1. Regard A as an SRSWOR of size n_A from P . Choose n_A units from A by a bootstrap method suitable for an SRSWOR of size n_A from P . Denote the sampled labels by A^* .
2. Regard B as an SRSWOR of size n_B from $P - A$ conditional on A having been selected. Choose n_B units from B by a bootstrap method suitable for an SRSWOR of size n_B from $P - A$. Denote the sampled labels by B^* .
3. For $j \in B^*$, define the mean-adjustment as \tilde{x}_j , where

$$\tilde{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_{A^*}) / (1 - f_A), \quad (1)$$

with $\bar{x}_A = n_A^{-1} \sum_{i \in A} x_i$, $\bar{x}_{A^*} = n_A^{-1} \sum_{i \in A^*} x_i$, and $f_A = n_A / N$.

4. Let $\mathbf{y}_{A^*} = \{y_i : i \in A^*\}$, $\mathbf{x}_{A^*} = \{x_i : i \in A^*\}$, and $\tilde{\mathbf{x}}_{B^*} = \{\tilde{x}_j : j \in B^*\}$. The bootstrap analogue of $\hat{\theta}$ is then given by $\hat{\theta}^* = t(\mathbf{y}_{A^*}, \mathbf{x}_{A^*}, \tilde{\mathbf{x}}_{B^*})$.

For bootstrap methods for a finite population, see Shao and Tu (1995, Chapter 6). The Bernoulli Bootstrap (BBE) proposed by Funaoka, Saigo, Sitter and Toida (2006) is appropriate for our method because of a reason specified later. To obtain a bootstrap sample A^* in the BBE, we conduct random replacement for each i in A : keep (x_i, y_i) in the bootstrap sample with probability $p = \{1 - (1 - n_A^{-1})^{-1} (1 - f_A)\}^{1/2}$ or replace it with one randomly selected from A . For the case where $p \notin [0, 1]$, see Funaoka *et al.* (2006).

To estimate the variance of $\hat{\theta}$, repeat steps 1-4 a large number of times K and use

$$v_{\text{boot}}(\hat{\theta}) = K^{-1} \sum_{k=1}^K (\hat{\theta}_{(k)}^* - \hat{\theta}_{(\cdot)}^*)^2, \quad (2)$$

where $\hat{\theta}_{(k)}^*$ is the value of $\hat{\theta}^*$ in the k^{th} bootstrap sample and $\hat{\theta}_{(\cdot)}^* = K^{-1} \sum_{k=1}^K \hat{\theta}_{(k)}^*$.

When f_A is negligible, the mean adjustment (1) is unnecessary. The above method then reduces for large n_A to that by Schreuder *et al.* (1987) and Sitter (1997).

The proposed bootstrap method is motivated by the following two observations. First, let sampling schemes I and II be $[P \rightarrow A + B, A + B \rightarrow A]$ and $[P \rightarrow A, P - A \rightarrow B]$, respectively, where \rightarrow means "the right hand side is an SRSWOR from the left hand side." Then, I and II

implement the identical design. In fact, the design probability assigned to a particular sample $\{\mathbf{i} = (i_1, i_2, \dots, i_{n_A}) \in A, \mathbf{j} = (j_1, j_2, \dots, j_{n_B}) \in B\}$ in I is $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_A+B} \times {}_{n_A+B} C_{n_A}]^{-1} = n_A! n_B! (N - n_{A+B})! / N!$ while it is $\Pr\{\mathbf{i} \in A, \mathbf{j} \in B\} = [{}_N C_{n_A} \times {}_{N-n_A} C_{n_B}]^{-1} = n_A! n_B! (N - n_{A+B})! / N!$ in II. Obviously, the sampling distribution of an estimator under repeated sampling depends on the sampling design. So, it is a matter of convenience to assume II is carried out even when I is employed.

Second, to motivate the mean adjustment (1), observe that the mean of x of the set $P - A$, or the conditional expectation of \bar{x}_B under repeated sampling given A , is $\bar{X}_{P-A} = (\bar{X} - f_A \bar{x}_A) / (1 - f_A)$. The bootstrap analogue of \bar{X}_{P-A} is given by $\bar{X}_{P-A^*} = (\bar{X} - f_A \bar{x}_{A^*}) / (1 - f_A)$. So, equation (1) amounts to $\tilde{x}_j = x_j - \bar{X}_{P-A} + \bar{X}_{P-A^*}$, a mean adjustment similar to that proposed by Rao and Shao (1992) in the context of hot deck imputation under the uniform response mechanism. This mean adjustment ensures appropriate correlations between x in A^* and x in B^* required for consistent variance estimation with high sampling fractions (see Rao and Sitter 1997, page 760). Note that the condition $n_A = n_{A^*}$ or $f_A = f_{A^*}$ is essential for cancelling out \bar{X} in the mean adjustment. Therefore, the mean-adjusted bootstrap requires a bootstrap method for SRSWOR which retains the original sample size, such as the BBE.

It is shown in Appendix A that the proposed bootstrap method provides design-consistent variance estimation for the class of estimators studied by Rao and Sitter (1997). Since no rescaling is performed, the method also works for estimation of distribution functions. Under some regularity conditions for the population distribution function, it provides design-consistent variance estimates for quantiles.

3. Illustrations

3.1 Ratio estimator

To illustrate, let us first consider the ratio estimator $\bar{y}_r = r_A \bar{x}_{A+B}$, where $r_A = \bar{y}_A / \bar{x}_A$, $w_A = n_A / n_{A+B}$, and $\bar{x}_{A+B} = w_A \bar{x}_A + (1 - w_A) \bar{x}_B$. Let $\bar{y}_r^* = (\bar{y}_{A^*} / \bar{x}_{A^*}) \{w_{A^*} \bar{x}_{A^*} + (1 - w_{A^*}) \tilde{x}_{B^*}\}$, the bootstrap analogue of \bar{y}_r . Using the results in Appendix A with $h(\bar{y}_A, \bar{x}_A, \bar{x}_B) = (\bar{y}_A / \bar{x}_A) \{w_A \bar{x}_A + (1 - w_A) \bar{x}_B\}$, we may approximate variance of \bar{y}_r^* under the proposed bootstrap method $V_*(\bar{y}_r^*)$ by

$$\begin{aligned} V_*(\bar{y}_r^*) &\doteq (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1 - f_A)}{n_A} \hat{S}_{dA}^2 \\ &+ 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1 - f_{A+B})}{n_{A+B}} r_A \hat{S}_{dA} \hat{S}_{dB} \\ &+ \frac{(1 - f_{A+B})}{n_{A+B}} r_A^2 \left[\frac{(w_A - f_A)}{(1 - f_A)} \hat{S}_{xA}^2 + \frac{(1 - w_A)}{(1 - f_A)} \hat{S}_{xB}^2 \right], \quad (3) \end{aligned}$$

where $\hat{S}_{dA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)^2$, $\hat{S}_{dA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - r_A x_i)(x_i - \bar{x}_A)$, $\hat{S}_{xA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^2$, and $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B)^2$. The right hand side of (3) can be described as a “bootstrap-linearization” variance estimator. We denote it by $v_{BL}(\bar{y}_r)$. Note that $v_{BL}(\bar{y}_r)$ is almost identical to the jackknife-linearization variance estimator by Rao and Sitter (1995),

$$\begin{aligned} v_{JL}(\bar{y}_r) &= (\bar{x}_{A+B} / \bar{x}_A)^2 \frac{(1-f_A)}{n_A} \hat{S}_{dA}^2 \\ &+ 2(\bar{x}_{A+B} / \bar{x}_A) \frac{(1-f_{A+B})}{n_{A+B}} r_A \hat{S}_{dA} \\ &+ \frac{(1-f_{A+B})}{n_{A+B}} r_A^2 \hat{S}_{xA}^2, \end{aligned} \quad (4)$$

where $\hat{S}_{xA+B}^2 = (n_{A+B} - 1)^{-1} \sum_{i \in A+B} (x_i - \bar{x}_{A+B})^2$, which agrees with equation 4.8 of Demnati and Rao (2004), page 25. Since they are close to $v_{JL}(\bar{y}_r)$, $V_*(\bar{y}_{lr})$, its Monte Carlo approximation $v_{boot}(\bar{y}_r)$ and $v_{BL}(\bar{y}_{lr})$ should perform well not only unconditionally but conditionally on $(\bar{x}_{A+B} / \bar{x}_A)$ as well. It is interesting to note that Taylor linearization in deriving $v_{BL}(\bar{y}_r)$ is performed around the sample means, not the population means (see the comment made by Demnati and Rao 2004, page 21).

3.2 Regression estimator

We next consider the regression estimator. The estimator of the population mean is $\bar{y}_{lr} = \bar{y}_A + b_A(\bar{x}_{A+B} - \bar{x}_A) = \bar{y}_A + (1-w_A)b_A(\bar{x}_B - \bar{x}_A)$, where $b_A = \hat{S}_{xA} / \hat{S}_{xA}^2$ with $\hat{S}_{xA} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)(y_i - \bar{y}_A)$. Let $\bar{y}_{lr} = \bar{y}_A + (1-w_A)b_A(\bar{x}_B - \bar{x}_A)$. Using the results in Appendix A (see also Appendix B), we have

$$\begin{aligned} V_*(\bar{y}_{lr}^*) &\doteq \frac{(1-f_A)}{n_A} m_{02} \\ &+ \frac{(1-f_{A+B})}{n_{A+B}} b_A^2 \left[\frac{(w_A - f_A)}{(1-f_A)} \hat{S}_{xA}^2 + \frac{(1-w_A)}{(1-f_A)} \hat{S}_{xB}^2 \right] \\ &+ z_A^2 \frac{(1-f_A)}{n_A} m_{22} + 2z_A \frac{(1-f_A)}{n_A} m_{12} \\ &+ 2z_A \frac{(1-f_{A+B})}{n_{A+B}} b_A m_{21} \\ &+ 4z_A^2 \frac{(1-f_A)}{n_A} a_A b_A \bar{x}_A \hat{S}_{xA}^2, \end{aligned} \quad (5)$$

where $z_A = n_A(\bar{x}_{A+B} - \bar{x}_A) / \{(n_A - 1) \hat{S}_{xA}^2\}$, $m_{pq} = (n_A - 1)^{-1} \sum_{i \in A} (x_i - \bar{x}_A)^p e_i^q$, $e_i = y_i - \bar{y}_A - b_A(x_i - \bar{x}_A)$, and $a_A = \bar{y}_A - b_A \bar{x}_A$. We call the right hand side of (5) a bootstrap-linearization variance estimator of \bar{y}_{lr} and denote it by $v_{BL}(\bar{y}_{lr})$. The jackknife-linearization variance estimator for \bar{y}_{lr} (Sitter 1997, page 781) is

$$\begin{aligned} v_{JL}(\bar{y}_{lr}) &= \frac{(1-f_A)}{n_A} m_{02} + \frac{(1-f_{A+B})}{n_{A+B}} b_A^2 \hat{S}_{xA+B}^2 \\ &+ \frac{z_A^2}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A)^2 e_i^2}{(1-c_i)^2} + \frac{2z_A}{n_A^2} \sum_{i \in A} \frac{(x_i - \bar{x}_A) e_i}{(1-c_i)} \\ &+ \frac{2z_A b_A}{n_A(n_{A+B} - 1)} \sum_{i \in A} \frac{(x_i - \bar{x}_A)(x_i - \bar{x}_{A+B}) e_i}{(1-c_i)}, \end{aligned} \quad (6)$$

where $c_i = n_A^{-1} + (x_i - \bar{x}_A)^2 / \{(n_A - 1) \hat{S}_{xA}^2\}$, the leverage values. From (5) and (6), $v_{boot}(\bar{y}_{lr})$, $v_{BL}(\bar{y}_{lr})$ and $v_{JL}(\bar{y}_{lr})$ perform in a similar fashion conditionally provided that $f_{A+B} \doteq 0$, n_A is large enough for all c_i to be nearly zero and the last term on the right hand side of (5) is negligible.

3.3 Estimation of distribution functions

As an example, let us take the model-calibrated pseudo-empirical maximum likelihood estimator (ME) under two-phase sampling proposed by Wu and Luan (2003) defined by

$$\hat{F}_{ME}(t) = \sum_{i \in A} \hat{p}_i I(y_i \leq t), \quad (7)$$

where \hat{p}_i maximizes the pseudo-likelihood function $\hat{l}(p) = \sum_A (N/n_A) \log p_i$ subject to (a) $\sum_A p_i = 1$ ($0 < p_i < 1$); and (b) $\sum_A p_i g_i = n_{A+B}^{-1} \sum_{i \in A+B} g_i$ where $g_i = g(x_i, t) = P(y \leq t | x_i)$ under a certain working model. For example, we may assume $\log(g_i / (1 - g_i)) = x_i' \theta$ with variance function $V(g) = g(1 - g)$. Chen, Sitter and Wu (2002) showed a simple algorithm for computing \hat{p}_i . It can be shown (see Wu and Luan 2003) that under the two-phase sampling considered in this paper,

$$\begin{aligned} \hat{F}_{ME}(t) &= n_A^{-1} \sum_{i \in A} I(y_i \leq t) \\ &+ \left\{ n_{A+B}^{-1} \sum_{i \in A+B} g_i - n_A^{-1} \sum_{i \in A} g_i \right\} \beta + o_p(n_A^{-1/2}), \end{aligned}$$

where $\beta = \sum_P (g_i - \bar{g}) I(y \leq t) / \sum_P (g_i - \bar{g})^2$ with $\bar{g} = N^{-1} \sum_P g_i$. Note that this equation is not used in estimation, but it shows that the variance of $\hat{F}_{ME}(t)$ can be estimated by the mean-adjusted bootstrap since $\hat{F}_{ME}(t)$ is approximated by a regression-type estimator.

3.4 Quantile estimation

Quantile estimation can be obtained by directly inverting $\hat{F}(t)$ by $\hat{F}^{-1}(\alpha) = \inf \{t : \hat{F}(t) \geq \alpha\}$ for some $\alpha \in (0, 1)$. For example, if (7) is used, then a quantile estimate is given by $y_{(k)}$, where $y_{(k)}$ is the k^{th} order statistic of y such that $\sum_{i=1}^{k-1} \hat{p}_{(i)} < \alpha$ and $\sum_{i=1}^k \hat{p}_{(i)} \geq \alpha$ (Chen and Wu 2002). Under some conditions specified in Chen and Wu (2002), a

Bahadur-type representation for $\hat{F}_{ME}^{-1}(\alpha)$ can be established. Thus the mean-adjusted bootstrap variance estimator for $\hat{F}_{ME}^{-1}(\alpha)$ is design-consistent. Note that no closed form variance estimator for $\hat{F}_{ME}^{-1}(\alpha)$ is available, but a consistent variance estimator based on Woodruff's interval estimation (Woodruff 1952) can be applied.

4. Simulation

4.1 Population and sampling

A simulation study was conducted to examine the mean-adjusted bootstrap variance estimator for the estimators in Section 3. We report here the results for estimating distribution functions and quantiles. The results for the ratio and regression estimators are available from the author upon request.

First, the auxiliary variable x for a finite population P of size $N = 2,000$ were generated as Gamma(1, 1). The characteristic variable y was then generated by $y_i = x_i + \sqrt{x_i} v_i$, where $v_i \sim N(0, 0.5^2)$. An SRSWOR $A + B$ of size $n_{A+B} = 800$ was taken from the population and then an SRSWOR A of size $n_A = 200$ was selected from $A + B$. The population was fixed throughout all simulation runs since we focus on design-based repeated-sampling properties.

4.2 Estimation of distribution functions

For the estimation of distribution functions, we took $\hat{F}_{ME}(t)$ as an example. Other estimators, e.g., Chambers and Dunstan (1986) and Rao, Kovar and Mantel (1990), can be handled similarly when an estimator is approximately design-unbiased. The working model for g in $\hat{F}_{ME}(t)$ was assumed to be logit with binomial variance. The bootstrap variance estimator $v_{boot}(\hat{F}_{ME}(t))$ was calculated with $K = 200$. The BBE was used in constructing a bootstrap sample. The total simulation runs were $M = 5,000$ while the true MSE of $\hat{F}_{ME}(t)$ at a given t was estimated by 50,000 runs.

We compared $v_{boot}(\hat{F}_{ME}(t))$ with three variance estimators: Wu and Luan's (2003) analytical estimator, the standard delete-1 jackknife and an *ad hoc* fpc-adjusted delete-1 jackknife. Wu and Luan's (2003) estimator is

$$v_a(\hat{F}_{ME}(t)) = (n_{A+B}^{-1} - N^{-1}) \hat{S}_I^2 + (n_A^{-1} - n_{A+B}^{-1}) \hat{S}_D^2,$$

where the two \hat{S}^2 components are estimated respectively by

$$\hat{S}^2 = s^2 + \left[\frac{1}{n_{A+B}(n_{A+B}-1)} \sum_{j>i; i, j \in A+B} u_{ij} - \frac{1}{n_A(n_A-1)} \sum_{j>i; i, j \in A} u_{ij} \right] \hat{\beta}_F,$$

where $s^2 = \{n_A(n_A-1)\}^{-1} \sum_{i<j; i, j \in A} v_{ij}^2$, and $\hat{\beta}_F = \sum_{i<j; i, j \in A} u_{ij} v_{ij} / \sum_{i<j; i, j \in A} u_{ij}^2$ with u_{ij} and v_{ij} specified as follows: For \hat{S}_I^2 , $v_{ij} = (I_j - I_j)^2$ and $u_{ij} = (\hat{g}_i - \hat{g}_j)^2$ with $I_j = I(y_j \leq t)$ and $\hat{g}_i = \hat{g}(x_i, t)$ estimated in A ; For \hat{S}_D^2 , $v_{ij} = (\hat{D}_i - \hat{D}_j)^2$ and $u_{ij} = \hat{g}_i(1 - \hat{g}_j) + \hat{g}_j(1 - \hat{g}_i)$ with $\hat{D}_i = I_i - \hat{g}_i \hat{\beta}$, $\hat{\beta} = \sum_{i \in A} I_i(\hat{g}_i - \bar{g}_A) / \sum_{i \in A} (\hat{g}_i - \bar{g}_A)^2$ and $\bar{g}_A = n_A^{-1} \sum_{i \in A} \hat{g}_i$.

The standard delete-1 jackknife formula is given by

$$v_J(\hat{\theta}) = \frac{(n_{A+B} - 1)}{n_{A+B}} \sum_{j \in A+B} (\hat{\theta}_{(-j)} - \hat{\theta}_{(.)})^2,$$

where $\hat{\theta} = \hat{F}_{ME}(t)$, $\hat{\theta}_{(-j)}$ is the j^{th} jackknife pseudo-estimate and $\hat{\theta}_{(.)} = n_{A+B}^{-1} \sum_{j \in A+B} \hat{\theta}_{(-j)}$. Note that for $j \in A$, both y_j and x_j are deleted from the sample while for $j \in B$, only x_j is deleted (see Rao and Sitter 1995 and Sitter 1997). The *ad hoc* fpc-adjusted formula is $v_{Jfpc}(\hat{F}_{ME}(t)) = (1 - f_{A+B}) v_J(\hat{F}_{ME}(t))$.

Table 1 shows the relative bias (%Bias) and the coefficient of variation (CV) of the four variance estimators for $\hat{F}_{ME}(t_{\alpha})$ ($\alpha = 0.10, 0.25, 0.50, 0.75, 0.90$), where $F(t_{\alpha}) = \alpha$. Here, %Bias and CV were calculated as %Bias = $100 \times (M^{-1} \sum_{m=1}^M v^{(m)} - \text{MSE}) / \text{MSE}$ and $\text{CV} = [M^{-1} \sum_{m=1}^M (v^{(m)} - \text{MSE})^2]^{1/2} / \text{MSE}$, respectively, where $v^{(m)}$ is a variance estimate in the m^{th} simulation run. Table 1 demonstrates that $v_J(\hat{F}_{ME}(t))$ is biased upward since the sampling fractions are not negligible, that $v_{Jfpc}(\hat{F}_{ME}(t))$ is biased downward since the *ad hoc* adjustment factor $(1 - f_{A+B})$ is too small, and that both $v_a(\hat{F}_{ME}(t))$ and $v_{boot}(\hat{F}_{ME}(t))$ are approximately unbiased although the latter is slightly more unstable, as is typical for a resampling method.

Table 1 Variance estimation for the pseudo-empirical MLE $\hat{F}_{ME}(t_{\alpha})$

		α				
Estimator		0.10	0.25	0.50	0.75	0.90
$v_{boot}(\hat{F}_{ME}(t_{\alpha}))$	%Bias	0.27	-0.22	0.64	0.83	2.73
	CV	0.19	0.14	0.14	0.15	0.24
$v_a(\hat{F}_{ME}(t_{\alpha}))$	%Bias	-2.29	-2.03	-0.47	-1.95	-3.26
	CV	0.17	0.11	0.09	0.11	0.19
$v_J(\hat{F}_{ME}(t_{\alpha}))$	%Bias	14.24	17.29	22.98	23.80	24.97
	CV	0.24	0.21	0.25	0.27	0.36
$v_{Jfpc}(\hat{F}_{ME}(t_{\alpha}))$	%Bias	-31.45	-29.63	-26.21	-25.72	-25.02
	CV	0.33	0.30	0.27	0.27	0.30

Paralleling Royall and Cumberland (1981a, 1981b), we ordered the $M = 5,000$ simulated samples on the values of $\bar{x}_{A+B} - \bar{x}_A$, classified them into 20 consecutive groups of $G = 250$ in each of which the simulated conditional MSE(MSE_c) and conditional mean of $v(E_c(v))$ were computed. Figure 1 shows MSE_c and $E_c(v)$ plotted against the group averages of $\bar{x}_{A+B} - \bar{x}_A$ for $t_{0.10}$ and $t_{0.90}$. It is seen that both $v_a(\hat{F}_{ME}(t))$ and $v_{boot}(\hat{F}_{ME}(t))$ behave

similarly conditioned on $\bar{x}_{A+B} - \bar{x}_A$. The jackknife variance estimators, $v_J(\hat{F}_{ME}(t))$ and $v_{Jpc}(\hat{F}_{ME}(t))$, though biased, track a trend in MSE_c .

4.3 Quantile estimation

By directly inverting $\hat{F}_{ME}(t)$, we estimated the α quantile. To obtain \hat{p}_i for $\hat{F}_{ME}(t)$, we fixed t at \hat{t}_α , where $\hat{t}_\alpha = \inf \{t: n_A^{-1} \sum_A I(y_i \leq t) \geq \alpha\}$, an estimator using only $\{y_i: i \in A\}$. For variance estimation, $K = 1,000$ bootstrap samples were created. For comparison, we also computed the Woodruff variance estimator (Woodruff 1952 and Shao and Tu 1995, page 238),

$$v_W(\hat{F}_{ME}^{-1}(\alpha)) = \left[\frac{\hat{F}_{ME}^{-1}(\alpha + \zeta_{1-\kappa/2} \hat{\sigma}_{\hat{F}}) - \hat{F}_{ME}^{-1}(\alpha - \zeta_{1-\kappa/2} \hat{\sigma}_{\hat{F}})}{2\zeta_{1-\kappa/2}} \right]^2,$$

where $\hat{\sigma}_{\hat{F}}^2 = v(\hat{F}_{ME}(t))$ with $t = \hat{F}_{ME}^{-1}(\alpha)$ and $\zeta_{1-\kappa/2}$ is the $(1 - \kappa/2)$ quantile of $N(0, 1)$. We let $\kappa = 0.05$ although the best choice of κ is unknown. The performance measures, %Bias and CV, were calculated through

$M = 5,000$ runs while the true MSE was estimated through 50,000 simulation runs.

Table 2 summarizes the results for quantile estimation. It demonstrates that the mean-adjusted bootstrap has an upward bias in estimating $V(\hat{F}_{ME}^{-1}(\alpha))$ while the bias in the Woodruff variance estimator is negligible.

Table 2 Variance estimation for quantiles

Estimator		α				
		0.10	0.25	0.50	0.75	0.90
$v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$	%Bias	6.27	14.32	10.05	10.02	10.28
	CV	0.53	0.53	0.51	0.52	0.61
$v_W(\hat{F}_{ME}^{-1}(\alpha))$	%Bias	1.64	3.75	2.92	0.70	-3.67
	CV	0.50	0.45	0.45	0.46	0.52

Figure 2 shows conditional properties of $v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$ and $v_W(\hat{F}_{ME}^{-1}(\alpha))$ for $\alpha = 0.10, 0.90$. We see that both $v_{boot}(\hat{F}_{ME}^{-1}(\alpha))$ and $v_W(\hat{F}_{ME}^{-1}(\alpha))$ track MSE_c similarly although the former uniformly possesses an upward bias.

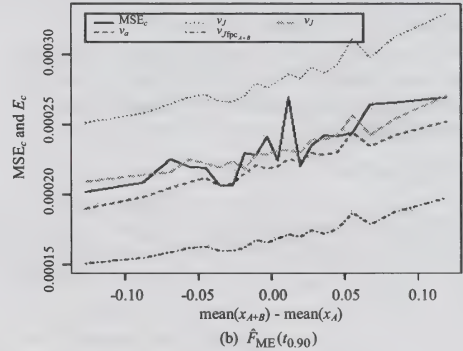
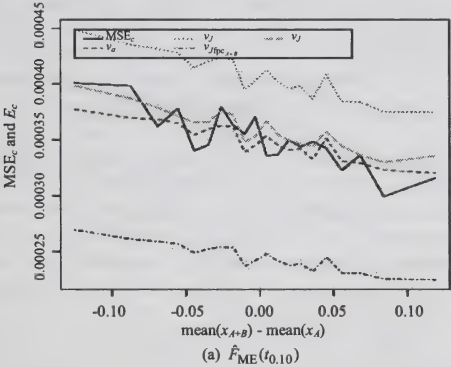


Figure 1 MSE_c and $E_c(v)$ for $\hat{F}_{ME}(t_\alpha)$

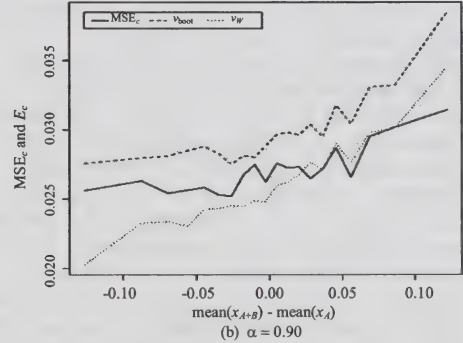
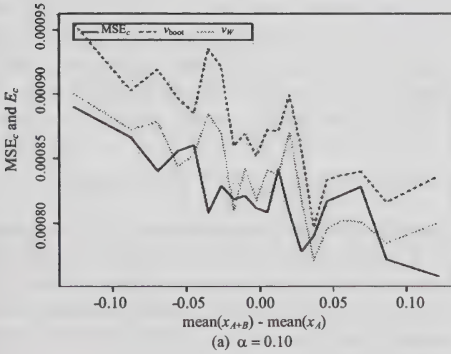


Figure 2 MSE_c and $E_c(v)$ for quantile estimation

5. Further remarks

5.1 Stratified two-phase sampling

Suppose a population is to be stratified into H strata but no information for stratification is available. A possible solution for this situation is to first obtain an SRSWOR of size n' from the population, observe auxiliary variables including the ones for stratification, stratify the sample into H strata, and in each stratum take an SRSWOR of size n_h from n'_h units belonging to stratum h in the sample. See, for example, Cochran (1977, section 12.2) for details.

Let N_h be the size of stratum h in the population. Conditioned on $n'_h > 0$, the first-phase sampling in stratum h described above is equivalent to simple random sampling without replacement of size n'_h in stratum h independent across strata. Thus, given n'_h ($h=1, \dots, H$), the mean-adjusted bootstrap can be applied independently in different strata to obtain a bootstrap sample. When N_h is unknown, as is usually the case for stratified two-phase sampling, an unbiased estimator $\hat{N}_h = N(n'_h/n')$ can be used in the mean-adjusted bootstrap. In this case, the sampling fraction n'/N is used commonly throughout all the strata.

Note, however, that the present discussion is legitimate for estimates conditioned on the first phase sample sizes. Variance due to the variable n'_h may be large. For unconditional variance estimation, see Kim *et al.* (2006).

5.2 Non-response

The above comment applies to imputed survey data under the uniform response mechanism. Let us suppose that a population is stratified into S_h ($h=1, \dots, H$) where simple random sampling without replacement is undertaken independently. A sample is divided into imputation classes C_l ($l=1, \dots, L$) in each of which the response rate is assume to be uniform and imputation is performed. An imputation class may cut across strata. We also assume which imputation class a sampled unit belongs to is correctly identified before imputation. Let us denote the numbers of sampled units and respondents in $S_h \cap C_l$ by n_{hl} and r_{hl} , respectively. Then, it is seen that given n_{hl} and r_{hl} , the corresponding design in $S_h \cap C_l$ is the same as the one discussed in this paper if we regard the n_{hl} units and r_{hl} respondents as $A+B$ and A , respectively. Therefore, the mean-adjusted bootstrap can be conducted independently in different $S_h \cap C_l$ ($h=1, \dots, H; l=1, \dots, L$). The size of $S_h \cap C_l$, denoted by N_{hl} , can be estimated by $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Note that this is a bootstrap method conditioned on the number of respondents.

6. Conclusion

In this paper, we have proposed the mean-adjusted bootstrap for two-phase sampling. The method requires a

simple mean adjustment and can handle the estimation of distribution functions and quantiles because it requires no rescaling. The Taylor series expansion shows that the method has desirable conditional properties for the ratio and regression estimators. A simulation study demonstrates that it also has similar conditional properties in estimating distribution functions and quantiles. An extension to stratified two-phase sampling is straightforward. Conditioned on the first phase sample sizes, the method can handle stratified two-phase sampling and imputation under the uniform response mechanism. We are currently investigating an extension of the proposed method to more generalized multi-phase sampling designs.

Acknowledgements

This research was supported by a grant from the Japan Society of the Promotion of Science. The author would like to thank Professor Randy R. Sitter, the Editor, the Associate Editor and the two referees for their helpful comments and suggestions.

Appendix A

In this appendix, we show that the proposed bootstrap method provides consistent variance estimates for a class of estimators considered by Rao and Sitter (1997). We use the same setting as in Rao and Sitter (1997) with slightly different notation. For simplicity, we assume there exists only one stratum, but an extension to stratified two-phase sampling is straightforward.

Consider a class of estimators, $\theta = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$, of a population parameter $\theta = h(\bar{Y}, \bar{X}, \bar{X})$, where \bar{Y} and \bar{X} are the population means of vectors y and x , i.e., $\bar{Y} = N^{-1} \sum_{i \in P} y_i$ and $\bar{X} = N^{-1} \sum_{i \in P} x_i$. Here, x is observed in the first phase sample $A+B$ whereas y is measured only in the second phase sample A . The sample means (\bar{y}_A, \bar{x}_A) and \bar{x}_B are calculated in A and B , respectively, i.e., $\bar{y}_A = n_A^{-1} \sum_{i \in A} y_i$, $\bar{x}_A = n_A^{-1} \sum_{i \in A} x_i$, and $\bar{x}_B = n_B^{-1} \sum_{i \in B} x_i$.

By a Taylor expansion, we have

$$\hat{\theta} = \theta + \nabla h'(\Delta \bar{y}_A', \Delta \bar{x}_A', \Delta \bar{x}_B')' + o_p(n_A^{-1/2}),$$

where ∇h is the gradient vector of h evaluated at $(\bar{Y}, \bar{X}, \bar{X})$, $\Delta \bar{y}_A = \bar{y}_A - \bar{Y}$, $\Delta \bar{x}_A = \bar{x}_A - \bar{X}$, $\Delta \bar{x}_B = \bar{x}_B - \bar{X}$, and $'$ means a transposed matrix (see equation 33.7 of Rao and Sitter 1997, page 757 and the required conditions therein). Then, the variance of $\hat{\theta} = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$ is approximated by

$$V(\hat{\theta}) \doteq \nabla h' \sum_{(\bar{y}_A', \bar{x}_A', \bar{x}_B')} \nabla h,$$

where $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$ is the variance-covariance matrix of $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ under repeated two-phase sampling. Because A and B are SRSWOR's of size n_A and n_B from the population P , respectively, we see that $\Sigma_{(\bar{y}_A, \bar{x}_A)'} = (1 - f_A) S_{y, x}^2 / n_A$ and $\Sigma_{\bar{x}_B} = (1 - f_B) S_x^2 / n_B$, where $S_u^2 = (N - 1)^{-1} \sum_{i \in P} (u_i - \bar{U})(u_i - \bar{U})'$ is the population variance of $u = (y, x)'$ or x and $f_B = n_B / N$. For $\text{Cov}(\bar{y}_A, \bar{x}_B)$, let E_A and $E_{B|A}$ be the expectation for selecting an SRSWOR A from P and choosing an SRSWOR B from $P - A$ given A , respectively. Note that $E_{B|A}(\bar{x}_B) = (\bar{X} - f_A \bar{x}_A) / (1 - f_A)$. So, we have

$$\begin{aligned} \text{Cov}(\bar{y}_A, \bar{x}_B) &= E(\bar{y}_A \bar{x}_B') - E(\bar{y}_A) E(\bar{x}_B') \\ &= E_A(\bar{y}_A E_{B|A}(\bar{x}_B)) - \bar{Y} \bar{X}' \\ &= -S_{yx} / N, \end{aligned}$$

where $S_{yx} = (N - 1)^{-1} \sum_{i \in P} (y_i - \bar{Y})(x_i - \bar{X})'$. Similarly, $\text{Cov}(\bar{x}_A, \bar{x}_B) = -S_{x^2}^2 / N$.

Now consider a Taylor expansion of $\hat{\theta}^* = h(\bar{y}_A, \bar{x}_A, \bar{x}_B')$ with $\bar{x}_B' = \bar{x}_B' + f_A(\bar{x}_A - \bar{x}_A') / (1 - f_A)$, the bootstrap analogue of $\hat{\theta} = h(\bar{y}_A, \bar{x}_A, \bar{x}_B)$. Let E_* and V_* be the expectation and variance under the proposed bootstrap procedure, respectively. First, observe that $E_*(\bar{y}_A) = \bar{y}_A$, $E_*(\bar{x}_A) = \bar{x}_A$ and

$$\begin{aligned} E_*(\bar{x}_B') &= E_{*A}(E_{*B|A}(\bar{x}_B')) \\ &= E_{*A}(\bar{x}_B + f_A(\bar{x}_A - \bar{x}_A') / (1 - f_A)) \\ &= \bar{x}_B, \end{aligned}$$

where E_{*A} and $E_{*B|A}$ are respectively the expectation with respect to sampling A^* and the conditional expectation with respect to sampling B^* given A^* under the proposed bootstrap method. Then, $\hat{\theta}^* = h(\bar{y}_A, \bar{x}_A, \bar{x}_B')$ is approximated by

$$\hat{\theta}^* = \hat{\theta} + \nabla h^* (\Delta \bar{y}_A', \Delta \bar{x}_A', \Delta \bar{x}_B')' + o_p(n_A^{-1/2}),$$

where ∇h^* is the gradient of h evaluated at $(\bar{y}_A, \bar{x}_A, \bar{x}_B)$, $\Delta \bar{y}_A' = \bar{y}_A' - \bar{y}_A$, $\Delta \bar{x}_A' = \bar{x}_A' - \bar{x}_A$, and $\Delta \bar{x}_B' = \bar{x}_B' - \bar{x}_B$ (see equation 33.A.1 of Rao and Sitter 1997, page 767 and the required conditions therein). Therefore, $V_*(\hat{\theta}^*)$ is approximated by

$$V_*(\hat{\theta}^*) \doteq \nabla h^* \Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^* \nabla h^*,$$

where $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$ is the variance-covariance matrix of $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ under the proposed bootstrap sampling.

Consistent variance estimation under the proposed method is proved by showing ∇h^* and $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$ are consistent for ∇h and $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$, respectively. Consistency of ∇h^* for ∇h follows from consistency of $(\bar{y}_A, \bar{x}_A, \bar{x}_B)'$ for $(\bar{Y}, \bar{X}, \bar{X})$ and continuity of h .

Consistency of $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$ can be shown as follows. First, since we use a bootstrap method suitable for simple random sampling without replacement in subsampling A^* , we have $\Sigma_{(\bar{y}_A, \bar{x}_A)'}^* = (1 - f_A) \hat{S}_{y, x}^2 / n_A$, where $\hat{S}_{uA}^2 = (n_A - 1)^{-1} \sum_{i \in A} (u_i - \bar{u}_A)(u_i - \bar{u}_A)'$ with $u = (y, x)'$. Second, because

1. $\Sigma_{\bar{x}_B'}^* = E_{*A}(V_{*B|A}(\bar{x}_B')) + V_{*A}(E_{*B|A}(\bar{x}_B'))$, where V_{*A} and $V_{*B|A}$ are respectively the variance with respect to sampling A^* and the conditional variance with respect to sampling B^* given A^* ,
2. $V_{*B|A}(\bar{x}_B') = (1 - f_{B|A}) \hat{S}_{xB}^2 / n_B$, where $\hat{S}_{xB}^2 = (n_B - 1)^{-1} \sum_{i \in B} (x_i - \bar{x}_B)(x_i - \bar{x}_B)'$ and $f_{B|A} = n_B / (N - n_A)$, and
3. $E_{*B|A}(\bar{x}_B') = \bar{x}_B + f_A(\bar{x}_A - \bar{x}_A') / (1 - f_A)$, we have $\Sigma_{\bar{x}_B'}^* = (1 - f_{B|A}) \hat{S}_{xB}^2 / n_B + f_A \hat{S}_{xA}^2 / (N - n_A)$. Since both \hat{S}_{xA}^2 and \hat{S}_{xB}^2 are consistent for S_x^2 , $\Sigma_{\bar{x}_B'}^*$ is consistent for $\Sigma_{\bar{x}_B} = (1 - f_B) S_x^2 / n_B$. Finally, we compute $\text{Cov}_*(\bar{y}_A, \bar{x}_B')$ and $\text{Cov}_*(\bar{x}_A, \bar{x}_B')$. For the former, we have

$$\begin{aligned} \text{Cov}_*(\bar{y}_A, \bar{x}_B') &= E_*(\bar{y}_A \bar{x}_B') - E_*(\bar{y}_A) E_*(\bar{x}_B') \\ &= E_{*A}(\bar{y}_A E_{*B|A}(\bar{x}_B')) - \bar{y}_A \bar{x}_B' \\ &= E_{*A}(\bar{y}_A \{\bar{x}_B + f_A(\bar{x}_A - \bar{x}_A') / (1 - f_A)\}) - \bar{y}_A \bar{x}_B' \\ &= -\hat{S}_{yxA} / N, \end{aligned}$$

where $\hat{S}_{yxA} = (n_A - 1)^{-1} \sum_{i \in A} (y_i - \bar{y}_A)(x_i - \bar{x}_A)'$. Similarly, $\text{Cov}_*(\bar{x}_A, \bar{x}_B') = -\hat{S}_{x^2}^2 / N$. This completes the proof of consistency of $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$ for $\Sigma_{(\bar{y}_A, \bar{x}_A, \bar{x}_B)'}^*$.

Appendix B

In this appendix, we derive $v_{BL}(\bar{y}_h)$. Under the mean-adjusted bootstrap,

$$\begin{aligned} \bar{y}_{lr}^* &= \bar{y}_A \\ &+ (1 - w_A) b_A \left\{ -\frac{(\bar{x}_A' - \bar{x}_A)}{(1 - f_A)} + (\bar{x}_B' - \bar{x}_B) + (\bar{x}_B - \bar{x}_A) \right\}. \end{aligned}$$

Define

$$\begin{aligned} \hat{\xi}_{pq}^* &= n_A^{-1} \sum_{i \in A} x_i^p y_i^q, \\ \hat{\xi}^* &= [\hat{\xi}_{10}^*, \hat{\xi}_{01}^*, \hat{\xi}_{11}^*, \hat{\xi}_{20}^*, \bar{x}_B']' \end{aligned}$$

and

$$\xi = [\bar{x}_A, \bar{y}_A, n_A^{-1} \sum_{i \in A} x_i y_i, n_A^{-1} \sum_{i \in A} x_i^2, \bar{x}_B] = E_*(\hat{\xi}^*).$$

Note that $b_A = (\hat{\xi}_{11}^* - \hat{\xi}_{10}^* \hat{\xi}_{01}^*) / (\hat{\xi}_{20}^* - \hat{\xi}_{10}^{*2})$. Let $\bar{y}_{lr}^* = h(\hat{\xi}^*)$. This expression is slightly different from that in Appendix A, but we may exploit independent subsampling of A^* and B^* . Then, by Taylor linearization of $\bar{y}_{lr}^* = h(\hat{\xi}^*)$ around ξ ,

we obtain $\bar{y}_{lr}^* \doteq \bar{y}_{lr} + \nabla h^*(\hat{\xi}^* - \xi)$ and $V_*(\bar{y}_{lr}^*) \doteq \nabla h^* \Sigma_{\hat{\xi}^*}^* \nabla h^{*T}$, where

$$\nabla h^* = [-b_A(1-w_A)/(1-f_A) - z_A(\bar{y}_A - 2b_A \bar{x}_A), 1 - z_A \bar{x}_A, z_A, -z_A b_A, b_A(1-w_A)]^T$$

and $\Sigma_{\hat{\xi}^*}^* = [v_{ij}]$ with

$$\begin{aligned} v_{11} &= c_A \hat{S}_{x_A}^2, \\ v_{21} &= c_A \hat{S}_{xy_A}, \\ v_{22} &= c_A \hat{S}_{y_A}^2, \\ v_{31} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(x_i - \bar{x}_A), \\ v_{32} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})(y_i - \bar{y}_A), \\ v_{33} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i y_i - \xi_{11})^2, \\ v_{41} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i - \bar{x}_A), \\ v_{42} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(y_i - \bar{y}_A), \\ v_{43} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})(x_i y_i - \xi_{11}), \\ v_{44} &= c_A (n_A - 1)^{-1} \sum_{i \in A} (x_i^2 - \xi_{20})^2, \\ v_{51} &= v_{52} = v_{53} = v_{54} = 0, \\ v_{55} &= \{n_B^{-1} - (N - n_A)^{-1}\} \hat{S}_{x_B}^2, \end{aligned}$$

$v_{ij} = v_{ji}$, and $c_A = (1 - f_A)/n_A$. Rewriting the moments from the origin as the central moments, noting that $y_i - \bar{y}_A = b_A(x_i - \bar{x}_A) + e_i$, and using properties of e_i as the least-squares residuals, we obtain the right hand side of (5) after some algebra.

References

- Berger, Y.G., and Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society*, B, 68, 531-547.
- Biemer, P.P., and Atkinson, D. (1993). Estimation of measurement bias using a model prediction approach. *Survey Methodology*, 19, 127-136.
- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.
- Chen, J., and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

- Cochran, W.G. (1977). *Sampling Techniques*. 3rd Edition. New York: John Wiley & Sons, Inc.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.
- Funaoka, F., Saigo, H., Sitter, R.R. and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32, 151-156.
- Kim, J.-K., Navarro, A. and Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.
- Lee, H., and Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.
- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.
- Rao, J.N.K., and Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. In *Survey Measurement and Process Quality: Wiley Series in Probability and Statistics*. (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York. 753-768.
- Royall, R.M., and Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.
- Royall, R.M., and Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Schreuder, H.T., Li, H.G. and Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag: New York.
- Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.
- Woodruff, R.S. (1952). Confidence intervals for median and other position measures. *Journal of the American Statistical Association*, 47, 635-646.
- Wu, C., and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.

On standard errors of model-based small-area estimators

Nicholas Tibor Longford¹

Abstract

We derive an estimator of the mean squared error (MSE) of the empirical Bayes and composite estimator of the local-area mean in the standard small-area setting. The MSE estimator is a composition of the established estimator based on the conditional expectation of the random deviation associated with the area and a naïve estimator of the design-based MSE. Its performance is assessed by simulations. Variants of this MSE estimator are explored and some extensions outlined.

Key Words: Composite estimation; Empirical Bayes estimation; Shrinkage; Small-area estimation.

1. Introduction

Design-based methods have over the years been proven to be inefficient for small-area estimation because, unlike empirical Bayes and related methods, they cannot make effective use of auxiliary information. However, the assumptions associated with the models that are applied remain a weakness of model-based methods because inferences based on them have the ubiquitous caveat of 'If the model is valid ...'. In the application of empirical Bayes models to small-area estimation, the local areas (districts) are associated with random effects. In the design-based perspective, this assumption is not valid because in a hypothetical replication of the survey the same districts would be realised (except for some districts that happen not to be represented in the sample drawn), and the target quantities associated with them would also be the same. That is, the districts should be associated with fixed effects. The lack of validity in this aspect of empirical Bayes models has no adverse impact on estimation of small-area quantities (means, totals, proportions, and the like). Associating small areas with random effects is key to borrowing strength from or exploiting the similarity of the areas, as well as to doing so across variables, time points, surveys and other data sources, but it distorts the assessment of the precision of the estimators. Some composite estimators and estimators of their mean squared errors have the same deficiency.

In the next section we diagnose this problem in detail, and in Section 3 propose a solution, which is then illustrated and assessed in Section 4 by simulations using a set of examples. These range from the simplest and most congenial (agreeing with most of the assumptions made) to more complex and realistic but least congenial, so as to explore the robustness of the method. Its fuller potential is discussed in the concluding section.

2. Fixed and random

By sampling variance of a general estimator $\hat{\theta}$ based on a given data-generating (sampling) process χ we understand the variation of the values of $\hat{\theta}(\mathbf{X})$ in replications of the processes that generate datasets \mathbf{X} and apply $\hat{\theta}$ to them. In the design-based perspective, the replication of a survey of a country with its division to D districts yields the same district-level population quantities $\theta_d, d = 1, \dots, D$; these D quantities are *fixed*. In contrast, each replication in the model-based perspective, using empirical Bayes models, starts by generating a fresh set of D values θ_d , independently of the previous replications.

We regard the design-based perspective as appropriate, because, in principle, each quantity θ_d could be established with precision and a hypothetical replication of the survey would draw a sample from the same population, with the same division of the country into its districts and the same values of the recorded variables for each member of the population. Most established design-based methods are valid when the survey is based on a perfect sampling frame, which contains no duplicates and is exclusive for the studied population, and the sampling design is implemented with perfection, without any departures from the protocol. That is, the estimators they yield are (approximately) unbiased, the expressions for their sampling variances are correct, or nearly so, and these variances are estimated with small or no bias.

In contrast, model-based methods carry a much heavier burden of assumptions that often cannot be verified. Various model diagnostic procedures are available, but they are all subject to uncertainty. Interpreting failure to find a contradiction as evidence of absence of any contradiction is a commonly committed logical inconsistency. It can be overcome only by quoting properties of estimators when the assumptions are not valid, but such methods are difficult to develop because of a wide range of model violations that

1. Nicholas Tibor Longford, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain. E-mail: NTL@SNTL.co.uk.

one would have to take into account. Yet, despite these drawbacks, model-based methods have proven their worth in small-area estimation and are nowadays rightly regarded as indispensable (Ghosh and Rao 1994; Rao 2003; and Longford 2005).

The EURAREA project (EURAREA Consortium 2004) carried out a large-scale simulation study involving sampling from artificially generated populations that resemble the human populations of several European countries and application of several classes of estimators. It confirmed the superiority of model-based estimators, with several qualifications, but reported rather disappointing results regarding estimators of their standard errors. We trace this problem to an averaging applied in deriving the standard errors of shrinkage estimators.

Suppose a population is divided into D districts, each of them of population size that can for all practical purposes be regarded as infinite, and independent simple random sampling schemes are applied in the districts. We assume that within each district d the outcome variable Y has the normal distribution with mean μ_d and the same variance σ_w^2 , $N(\mu_d, \sigma_w^2)$. For the within-district population means μ_d , we assume the superpopulation model $\mu_d \sim N(\mu^*, \sigma_B^2)$, but we want to make inferences about a fixed set of (realised) means $\{\mu_d\}$. In Section 5, we discuss the more general regression setting defined by the within-district models

$$(Y|d) \sim N(\mathbf{X}_d\boldsymbol{\beta} + \delta_d, \sigma_w^2),$$

in which \mathbf{X}_d are the within-district regression matrices, $\boldsymbol{\beta}$ the set of corresponding regression parameters common to the districts, and δ_d is the deviation of the within-district regression from the typical regression defined by $\delta_d = 0$. In the superpopulation, δ_d are a random sample from $N(0, \sigma_B^2)$, but we want to make inferences about the fixed (realised) set $\{\delta_d\}$. Thus, we use model-based estimators, but assess their properties by design-based criteria.

Denote by μ the (national) mean of the quantities μ_d and by σ_B^2 the district-level variance, $\sigma_B^2 = D^{-1} \sum_d (\mu_d - \mu)^2$. Note that they differ from their respective superpopulation counterparts μ^* and σ_B^{*2} . We assume first that σ_B^2 , σ_w^2 and μ are known. Let $\hat{\mu}_d$ and $\hat{\mu}$ be the sample means of the variable of interest in district d and in the whole domain (country). They are based on samples of respective sizes n_d and $n = n_1 + \dots + n_D$. When no covariates are used the empirical Bayes (shrinkage) estimator of μ_d is

$$\tilde{\mu}_d = \left(1 - \frac{1}{1 + n_d\omega}\right) \hat{\mu}_d + \frac{1}{1 + n_d\omega} \hat{\mu}, \quad (1)$$

where $\omega = \sigma_B^2/\sigma_w^2$ is the variance ratio. The model-based conditional variance of μ_d , given the data, μ , σ_w^2 and σ_B^2 , equal to $\sigma_B^2/(1 + n_d\omega)$, is often regarded as the sampling

variance of $\tilde{\mu}_d$; the origins of this practice can be traced to the application of the EM algorithm. A more careful derivation acknowledges that in the design-based perspective $\tilde{\mu}_d$ is biased for μ_d ,

$$E(\tilde{\mu}_d|\mu_d) - \mu_d = -\frac{\mu_d - \mu}{1 + n_d\omega},$$

and its mean squared error is

$$\begin{aligned} \text{MSE}(\tilde{\mu}_d; \mu_d) &= \left(1 - \frac{1}{1 + n_d\omega}\right)^2 \text{var}(\hat{\mu}_d) + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2} \\ &= \sigma_w^2 \frac{n_d\omega^2}{(1 + n_d\omega)^2} + \frac{(\mu_d - \mu)^2}{(1 + n_d\omega)^2}, \end{aligned} \quad (2)$$

assuming, for simplicity, that $\hat{\mu} \equiv \mu$. To emphasise that MSE depends on the target, we include both the estimator and the target in its argument. In particular, $\text{MSE}(\tilde{\mu}; \mu) \neq \text{MSE}(\tilde{\mu}; \mu_d)$, unless $\mu_d = \mu$. An inconvenient feature of the identity in (2) is that it involves μ_d , the target of estimation. If we replace $(\mu_d - \mu)^2$ with its expectation over the districts, σ_B^2 , we obtain the more familiar identity

$$\overline{\text{MSE}}(\tilde{\mu}_d; \mu_d) = \frac{\sigma_B^2}{1 + n_d\omega}, \quad (3)$$

the EM-related conditional model-based variance of μ_d . The bar over MSE indicates expectation (averaging) of $(\mu_d - \mu)^2$, the numerator in the last term of (2), over the districts, with the sample sizes n_d intact. Throughout, we condition on the within-district sample sizes n_d , $d = 1, \dots, D$, even though in the sampling design each of them may be variable. $\overline{\text{MSE}}$ can be interpreted as model expectation, although the expectation or average of the squared deviations $(\mu_d - \mu)^2$ could be considered and estimated for a given set of districts without any reference to a model. The conditional variance in (3) is appropriate for districts with μ_d in the 'typical' distance, σ_B , from the national mean μ . When $|\mu_d - \mu| \neq \sigma_B$, an unbiased estimator of the conditional variance $\sigma_B^2/(1 + n_d\omega)$ is biased for $\text{MSE}(\tilde{\mu}_d; \mu_d)$. As the bias is related to the population quantity $\mu_d - \mu$, it is not reduced by increasing the sample size n_d .

3. Composite estimation of MSE

To estimate $\text{MSE}(\tilde{\mu}_d; \mu_d)$, we reuse the idea of shrinkage and combine the alternative estimators, $\sigma_B^2/(1 + n_d\omega)$ and a naïve estimator of the MSE in (2). This composite estimator can be motivated as follows. If $n_d = 0$, and therefore $\tilde{\mu}_d = \hat{\mu}$, we have no direct information about μ_d , so we cannot improve on $\sigma_B^2/(1 + n_d\omega)$ as an estimator of $\text{MSE}(\tilde{\mu}_d; \mu_d)$. When n_d is large, μ_d is estimated with precision sufficient for using $(\tilde{\mu}_d - \hat{\mu})^2$, possibly with an adjustment for bias, as an estimator of $(\mu_d - \mu)^2$. For

intermediate sample sizes, we search for a composition (compromise) of these two alternatives that are suitable in the extreme settings, when $n_d = 0$ and as $n_d \rightarrow +\infty$. We therefore derive expressions for their MSEs and then for the MSE of their combination.

We regard the constant $\sigma_B^2/(1+n_d\omega)$ as an estimator, and refer to it as the *averaged* estimator of MSE. Although it has no variance, it is biased, with mean squared error

$$\begin{aligned} \text{MSE} \left\{ \frac{\sigma_B^2}{1+n_d\omega}; \text{MSE}(\hat{\mu}_d; \mu_d) \right\} \\ = \left\{ \frac{\sigma_B^2}{1+n_d\omega} - \frac{\sigma_W^2 n_d \omega^2}{(1+n_d\omega)^2} - \frac{(\mu_d - \mu)^2}{(1+n_d\omega)^2} \right\}^2 \\ = \left\{ \frac{\sigma_B^2 - (\mu_d - \mu)^2}{(1+n_d\omega)^2} \right\}^2. \end{aligned} \quad (4)$$

The squared deviation $(\mu_d - \mu)^2$, involved in (2), is estimated naïvely by $(\hat{\mu}_d - \hat{\mu})^2$ with bias equal to $\sigma_W^2(n_d^{-1} - n^{-1}) \doteq \sigma_W^2/n_d$ and, assuming that $\hat{\mu}_d$ is normally distributed,

$$\begin{aligned} \text{MSE}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \\ = \text{var}\{(\hat{\mu}_d - \hat{\mu})^2 | \mu_d\} \\ + [E\{(\hat{\mu}_d - \hat{\mu})^2 - (\mu_d - \mu)^2 | \mu_d\}]^2 \\ \doteq \frac{2\sigma_W^4}{n_d^2} + 4(\mu_d - \mu)^2 \frac{\sigma_W^2}{n_d} + \frac{\sigma_W^4}{n_d^2} \\ = \frac{\sigma_W^2}{n_d} \left\{ \frac{3\sigma_W^2}{n_d} + 4(\mu_d - \mu)^2 \right\}, \end{aligned} \quad (5)$$

derived from the properties of the non-central χ^2 distribution and an approximation by letting $n \rightarrow +\infty$. As an alternative, $\hat{\mu}_d$ may be used instead of $\hat{\mu}_d$; elementary operations yield the approximations

$$\begin{aligned} E\{(\hat{\mu}_d - \hat{\mu})^2 | \mu_d\} &\doteq (1-b_d)^2 \left\{ \frac{\sigma_W^2}{n_d} + (\mu_d - \mu)^2 \right\} \\ \text{var}\{(\hat{\mu}_d - \hat{\mu})^2 | \mu_d\} &\doteq \frac{(1-b_d)^4}{n_d^2} \sigma_W^2 \{2\sigma_W^2 + 4n_d(\mu_d - \mu)^2\}, \\ \text{where } b_d &= 1/(1+n_d\omega), \text{ and so} \\ \text{MSE}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \\ &= \text{var}\{(\hat{\mu}_d - \hat{\mu})^2 | \mu_d\} + [E\{(\hat{\mu}_d - \hat{\mu})^2 - (\mu_d - \mu)^2 | \mu_d\}]^2 \\ &\doteq (1-b_d)^4 \frac{3\sigma_W^4}{n_d^2} \\ &\quad + 2(1-b_d)^2 (2-6b_d+3b_d^2) \frac{\sigma_W^2(\mu_d - \mu)^2}{n_d} \\ &\quad + b_d^2 (2-b_d)^2 (\mu_d - \mu)^4. \end{aligned} \quad (6)$$

This approximation is valid only for $b_d = 1/(1+n_d\omega)$, so further approximation is involved when we substitute a possibly suboptimal choice or an estimate of b_d based on an estimate of ω . In general, the coefficient b_d that minimises the MSE in (6) differs from $1/(1+n_d\omega)$ because the shrinkage with $b_d = 1/(1+n_d\omega)$ is optimal only for targets that are linear transformations of μ_d (Shen and Louis 1998). We do not pursue this avenue because the solution, being a complicated function of the parameters, is likely to be sensitive to the error in estimation of the parameters. The estimator $(\hat{\mu}_d - \hat{\mu})^2$ could be corrected for its bias in estimating $(\mu_d - \mu)^2$, although this may result in a negative estimate, especially when n_d is small.

Finally, we combine the two (biased) estimators of $\text{MSE}(\hat{\mu}_d; \mu_d)$, the averaged estimator $\sigma_B^2/(1+n_d\omega)$ and the naïve estimator derived from the identity in (2), using $(\hat{\mu}_d - \hat{\mu})^2$ as an estimator of $(\mu_d - \mu)^2$. The MSEs of these two estimators depend on $(\mu_d - \mu)^2$, so we replace the relevant terms by their expectations across the districts d . We replace $(\mu_d - \mu)^2$ with σ_B^2 , and $(\mu_d - \mu)^4$ with $3\sigma_B^4$ or, in general, with $\kappa\sigma_B^4$, where κ is the kurtosis of the (district-level) distribution of μ_d . Although it may at first appear that we have not gained anything, because we still have to remove the dependence of MSE on $(\mu_d - \mu)^2$ by using σ_B^2 instead, now we make this step at a later stage. In the simulations in Section 4, we show that this reduces the undesirable impact of averaging.

Thus, we search for the coefficient c_d that minimises the expected MSE of the composite estimator of the MSE,

$$\begin{aligned} \overline{\text{MSE}}(\hat{\mu}_d; \mu_d) \\ = (1-c_d) \overline{\text{MSE}}(\hat{\mu}_d; \mu_d) + c_d \overline{\text{MSE}}(\hat{\mu}_d; \mu_d) \\ = (1-c_d) \left\{ (1-b_d)^2 \frac{\sigma_W^2}{n_d} + b_d^2 (\hat{\mu}_d - \hat{\mu})^2 \right\} + c_d b_d \sigma_B^2. \end{aligned} \quad (7)$$

To evaluate the MSE of this MSE estimator, as a function of c_d , we use the expressions

$$\begin{aligned} \overline{\text{MSE}}\{b_d \sigma_B^2; \text{MSE}(\hat{\mu}_d; \mu_d)\} &= 2b_d^4 \sigma_B^4, \\ \overline{\text{MSE}}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} &\doteq \frac{\sigma_W^4}{n_d^2} (3+4n_d\omega), \\ \overline{\text{MSE}}\{(\hat{\mu}_d - \hat{\mu})^2; (\mu_d - \mu)^2\} \\ &\doteq \frac{\sigma_W^4}{n_d^2} \{3(1-b_d)^4 + 3b_d^2(2-b_d)^2 n_d^2 \omega^2 \\ &\quad + 2(1-b_d)^2 (2-6b_d+3b_d^2) n_d \omega\}, \end{aligned}$$

derived by averaging of the respective equations (4), (5) and (6); $(\mu_d - \mu)^2$ is replaced by σ_B^2 and $(\mu_d - \mu)^4$ by $3\sigma_B^4$.

Assuming that the district-level targets μ_d are normally distributed, the MSE of the composite estimator in (7) is

$$\begin{aligned}
& E \left\{ (1 - c_d)(1 - b_d)^2 \frac{\sigma_w^2}{n_d} + (1 - c_d)b_d^2(\hat{\mu}_d - \hat{\mu})^2 \right. \\
& \quad + c_d b_d \sigma_B^2 - b_d^2 \sigma_w^2 n_d \omega^2 - b_d^2 (\mu_d - \mu_d)^2 \}^2 \\
& = b_d^4 E \left\{ (1 - c_d) \sigma_B^2 n_d \omega + (1 - c_d)(\hat{\mu}_d - \hat{\mu})^2 \right. \\
& \quad \left. + c_d \sigma_B^2 (1 + n_d \omega) - \sigma_B^2 n_d \omega - (\mu_d - \mu)^2 \right\}^2 \\
& = b_d^4 E \left\{ (1 - c_d)(\hat{\mu}_d - \hat{\mu})^2 + c_d \sigma_B^2 - (\mu_d - \mu)^2 \right\}^2 \\
& \doteq b_d^4 \left[(1 - c_d)^2 \left(\frac{2\sigma_w^4}{n_d^2} + \frac{4\sigma_w^2}{n_d} (\mu_d - \mu)^2 \right) \right] \\
& \quad + b_d^4 \left[(1 - c_d) \frac{\sigma_w^2}{n_d} + c_d \{ \sigma_B^2 - (\mu_d - \mu)^2 \} \right]^2,
\end{aligned}$$

using the identities $(1 - b_d)^2 = b_d^2 n_d^2 \omega^2$ and $\sigma_B^2 = \sigma_w^2 \omega$ to extract the factor b_d^4 . By taking the expectation over the districts, keeping the sample sizes intact, we obtain

$$\begin{aligned}
& \overline{\text{MSE}} \{ \text{MSE}(\hat{\mu}_d; \mu_d) \} \\
& \doteq \frac{b_d^4}{n_d^2} \{ (1 - c_d)^2 (3 + 4n_d \omega) \sigma_w^4 + 2c_d^2 n_d^2 \sigma_B^4 \}.
\end{aligned}$$

The minimum of this quadratic function of c_d is attained for

$$c_d^* = \frac{3 + 4n_d \omega}{3 + 4n_d \omega + 2n_d^2 \omega^2}.$$

This choice of a coefficient c_d agrees with our expectations. For $n_d = 0$, $c_d^* = 1$ and we rely solely on the averaged MSE estimator, equal to σ_B^2 . Further, c_d^* is a decreasing function of n_d , converging to zero as n_d diverges to $+\infty$; for large n_d we rely on the naïve estimator of MSE. It is also a decreasing function of ω ; for $\omega = 0$, that is, $\sigma_B^2 = 0$, $c_d^* = 1$ for every district d , confirming that $\mu_d \equiv \mu$ and μ_d would be estimated precisely if μ were known. With increasing ω , $\sigma_B^2/(1 + n_d \omega)$ becomes less and less useful because the squared deviations $(\mu_d - \mu)^2$ are widely spread (around σ_B^2).

If we adjust $(\hat{\mu}_d - \hat{\mu})^2$ for its bias in estimating $(\mu_d - \mu)^2$, the expected MSE of the shrinkage estimator is minimised for

$$c_d^\dagger = \frac{1 + 2n_d \omega}{(1 + n_d \omega)^2}.$$

It is easy to check that

$$c_d^* - c_d^\dagger = \frac{n_d^2 \omega^2}{(1 + n_d \omega)^2} \frac{1}{3 + 4n_d \omega + 2n_d^2 \omega^2},$$

so the bias-adjusted estimator derived from (2) is assigned greater weight (equal to $1 - c_d^\dagger$) than the naïve estimator would be. But the difference is small for all values of $n_d \omega$.

The composite MSE estimator based on $(\bar{\mu}_d - \hat{\mu})^2$ is derived similarly, but the resulting expression is much more complex. The optimal shrinkage coefficient is

$$\begin{aligned}
c_d^{**} & = 3(1 - b_d)^4 + 2(1 - b_d)^2 f(b_d) n_d \omega - b_d (2 - b_d) f(b_d) n_d^2 \omega^2 \\
& \quad \times [3(1 - b_d)^4 + 2(1 - b_d)^2 f(b_d) n_d \omega - \\
& \quad \{2 - 4b_d(2 - b_d) + 3b_d^2 f(b_d)\} n_d^2 \omega^2],
\end{aligned}$$

where $f(b_d) = 2 - 6b_d + 3b_d^2$. The dependence on b_d is particularly problematic, because in practice b_d is estimated and the properties of the MSE estimator based on estimated c_d^{**} are bound to be affected by the uncertainty about b_d . In the derivations, we used the identity $b_d = 1/(1 + n_d \omega)$, so this expression could not be used when the values of b_d are set *a priori*.

4. Simulations

Properties of the composite estimator of MSE cannot be derived analytically, and so we resort to simulations. We consider the artificial setting of a national survey with a stratified sampling design, with strata coinciding with the country's 100 districts for which estimates of the means of a variable Y are sought. Simple random sampling is applied within each stratum, assumed to be of practically infinite population size. We have generated the values of the means μ_d from the normal distribution $N(\mu = 20, \sigma_B^2 = 8)$, and the sample sizes n_d from scaled conditional beta distributions, given the means μ_d , so as to inject a modicum of dependence of the means on the sample sizes. With this adjustment, the assumption underlying the averaged MSE estimator is false, but this could not be detected by a diagnostic procedure or a hypothesis test, not even with μ_d known. The sample size of one district was altered to be much greater than the rest, to represent the capital of the fictitious country. The within-stratum distributions of Y are $N(\mu_d, \sigma_w^2 = 100)$. The district-level means and sample sizes are fixed in the replications. For orientation, they are plotted in Figure 1. The districts are assigned order numbers from 1 to 100 in the ascending order of their sample sizes. The smallest sample size is $n_1 = 15$ and the overall sample size is 3,698.

In the simulations, comprising 1,000 replications, we generate the direct estimates $\hat{\mu}_d$ as independent random draws from $N(\mu_d, \sigma_w^2/n_d)$ and the within-district corrected sums of squares as independent draws from the appropriately scaled χ^2 distributions with $n_d - 1$ degrees of freedom. Then we evaluate the shrinkage estimator $\bar{\mu}_d$ for each district d , followed by evaluation of the averaged, naïve and the two composite MSE estimators using the coefficients c_d^* and c_d^\dagger or their naïve estimates.

In the first set of replications, we assume that μ , σ_w^2 and σ_B^2 are known, so that the simulation reproduces the theoretically derived results and enables us to assess the quality of the composite MSE estimators without the interference of uncertainty about the shrinkage coefficient $b_d = 1/(1 + n_d\omega)$. The results are summarised graphically in Figure 2. The empirical biases (their absolute values) of the four MSE estimators are plotted in the left-hand panel. Circles and black dots are used for the averaged and naïve estimators, respectively, and the biases of the composite estimators are connected by solid lines. The absolute values of the empirical biases are plotted, to highlight their strong association with the sample size for the naïve estimator. For

60 districts (60%), the composite estimator of MSE has a positive bias. For the naïve estimator, this count or percentage is higher (78), and for the averaged estimator lower (52). Throughout, the main contributor to the bias of the averaged MSE estimator is the deviation of the squared distance $(\mu_d - \mu)^2$ from the district-level variance σ_B^2 . The two composite estimators, based on $(\hat{\mu}_d - \hat{\mu})^2$ and on its bias-adjusted version, differ so little that their biases cannot be distinguished in the plot. The diagram shows that the averaged estimator of MSE entails substantial bias for a few districts, including several with large sample sizes. The biases of the naïve and composite estimators are without such extremes.



Figure 1 The district-level sample sizes and population means of *Y*. Artificially generated values

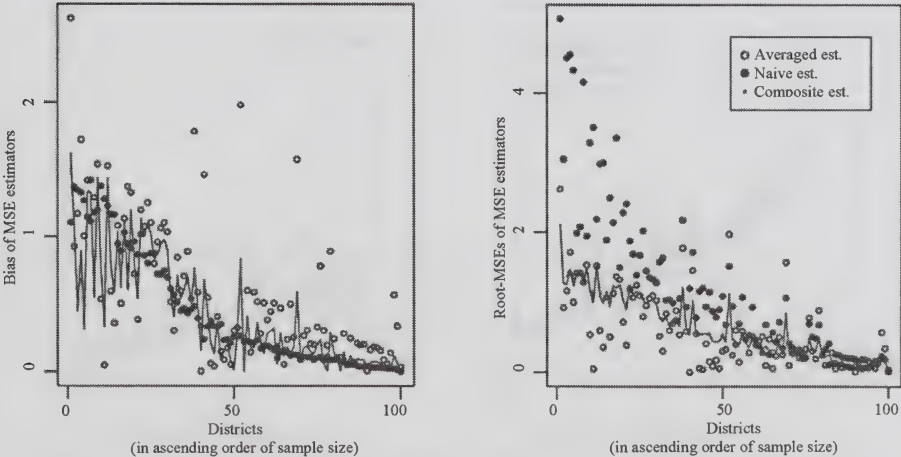


Figure 2 The bias and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. Based on simulations with an artificial setting. The bias and root-MSE of the composite estimators are connected by solid lines

In the right-hand panel, the root-MSEs of the MSE estimators are plotted, using the same symbols and layout. The diagram shows that the naïve estimator is inefficient, especially for districts with the smallest sample sizes, whereas the averaged estimator is very efficient for some but inefficient for some other districts, without any apparent relation to their sample sizes. In fact, apart from sample size, high efficiency is associated with proximity of $(\mu_d - \mu)^2$ to σ_B^2 and low efficiency with the smallest and largest values of $(\mu_d - \mu)^2$. For example, the empirical root-MSE of the averaged MSE estimator for district 1, with $n_1 = 15$, is 2.63, whereas its counterpart for district 11 ($n_{11} = 16$) is 0.049. Their population means are $\mu_1 = 24.55$, exceeding $\mu + \sigma_B$ by 1.72, and $\mu_{11} = 22.87$, differing from $\mu + \sigma_B$ by only 0.04. The root-MSEs of the naïve estimator are 5.08 and 3.51, and those of the composite estimator are 2.10 and 1.00 for the respective districts 1 and 11. The composite MSE estimator performs much more evenly, moderating the deficiencies of the averaged and naïve estimators.

All three estimators are conservative (have positive biases) for districts with relatively small MSE of $\hat{\mu}_d$. The averaged estimator has negative biases when the MSEs are relatively large. The composite estimator also has negative biases for such districts, but they tend to be smaller in absolute value. For districts with the smallest sample sizes, the composite estimator is not very effective because the naïve estimator is very inefficient. For a few of these districts, the composition is counterproductive, as a result of averaging, but such districts cannot be identified from a single realisation of the survey.

Next we study a less congenial setting, in which the normality assumptions of μ_d across the districts and of the

elementary observations y_{dj} within the districts are still satisfied, but the global parameters, μ , σ_W^2 and σ_B^2 , are not known and are estimated. We use the same means μ_d and sample sizes n_d as in Figure 1. The results of the simulations are summarised in Figure 3. In the left-hand panel, the empirical means of the MSE estimators are plotted, using the same symbols as in Figure 2, together with the empirical MSEs (crosses '+') of the shrinkage estimators $\hat{\mu}_d$. The empirical means of the averaged estimators have a regular pattern because the estimates in each replication depend only on the sample size n_d and the estimated variance ratio $\hat{\omega}$. For biases, the naïve estimators have a regular pattern, similar to their pattern in Figure 2. The naïve estimators have positive biases that decline with the sample size. The averaged estimators are far too conservative; their means do not veer from the smooth trend. The composite MSE estimators deviate from this trend in the appropriate direction, but not to full degree. Their average bias is positive, equal to 0.22, or 10% (2.42 vs. 2.20), and they overestimate the target MSE for 70 out of the 100 districts.

The right-hand panel displays the root-MSEs of the MSE estimators. The naïve estimator is inefficient, whereas the averaged estimator is very efficient for some and rather inefficient for other districts. The composite MSE estimator is more efficient than either naïve or averaged estimator for 36 districts; it is more efficient than the averaged estimator in exactly half of the districts, but it does not have its glaring weaknesses. As in the congenial setting (Figure 2), the differences due to bias adjustment of $(\hat{\mu}_d - \hat{\mu})^2$ in composite MSE estimation (using coefficients c_d^* or c_d^{**}) are negligible.

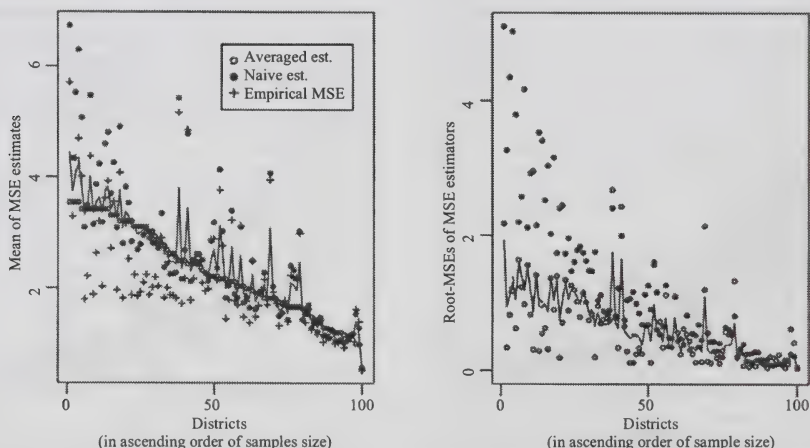


Figure 3 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators. The global parameters μ , σ_W^2 and σ_B^2 are estimated

Next we compare the MSE estimators for the district-level means of $Y^2/100$, denoted by v_d . The assumptions of normality both within and across districts are no longer appropriate. We apply the methods that rely on the normality assumptions, to assess the robustness of the composite estimators, but also to contrast the deficiencies of the averaging with the consequences of using 'incorrect' models. We chose the square transformation because the within-district expectations are known, equal to $(\mu_d^2 + \sigma_w^2)/100$, and could be estimated by

$$\hat{v}_d^* = \frac{\hat{\mu}_d^2 - \widehat{\text{MSE}}(\hat{\mu}_d) + \hat{\sigma}_w^2}{100}. \quad (8)$$

We denote by \hat{v}_d the empirical Bayes estimators applied to $y_d^2/100$.

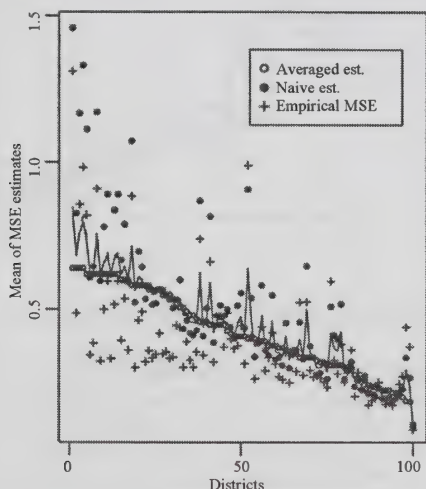
The results of the simulations based on the values of $y_d^2/100$ are presented in Figure 4, using the same layout and symbols as in Figure 3. The same conclusions about the biases and root-MSEs are arrived at as before, except that the naive estimator is even more inefficient and the performance of the averaged estimator even more erratic - it is both very efficient and inefficient for more districts than in the more congenial setting of Figure 3. The naive estimator is conservative, but for some districts with small n_d far too much so, and its MSEs for these districts are very large.

We contrast these conclusions with a comparison of estimating the district-level means of $Y^2/100$ by \hat{v}_d^* , transforming the estimates $\hat{\mu}_d$ according to (8). The estimator \hat{v}_d^* is more efficient than \hat{v}_d for most districts (90, in fact), and when less efficient, the relative difference of their MSEs is less than 4%. For a few districts, the

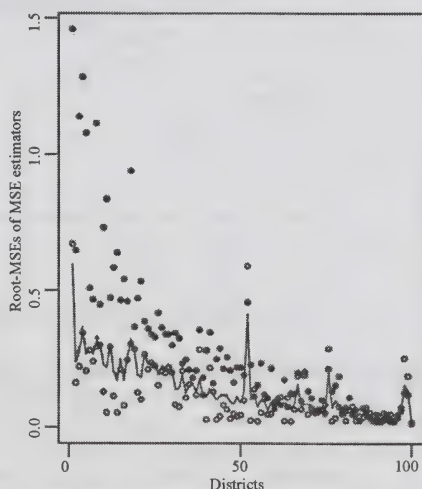
difference in efficiency is perceptible, exceeding 20% for ten districts. However, the differences in the MSEs are small in comparison with the biases in estimating these MSEs, as shown in Figure 5. The biases and MSEs of \hat{v}_d are marked by black dots connected to their counterparts for \hat{v}_d^* .

Part of the lack of efficiency of \hat{v}_d is due to its bias; the bias of \hat{v}_d exceeds the bias of \hat{v}_d^* for all but two districts, but the difference is non-trivial only when both estimators are positively biased. Thus, little efficiency is gained by arranging the analysis so that the distributional assumptions are satisfied. The gains are modest in comparison with the increase in the difficulty of estimating the efficiency, as expressed by $\text{MSE}(\hat{v}_d^*; v_d)$. Although the sampling variation of $\hat{\sigma}_w^2$ is trivial in large-scale surveys, the contribution of $\text{MSE}(\hat{\mu}_d; \mu_d)$ to $\text{MSE}(\hat{v}_d^*; v_d)$ cannot be ignored.

Figure 6 compares the composite MSE estimator with the naive estimator of MSE of $\hat{\mu}_d$ based on the empirical Bayes estimator of μ_d ; it is derived by substituting $\hat{\mu}_d$ for μ_d in (2). For brevity, we refer to it as the EB-naïve estimator. As anticipated in Section 3, it tends to underestimate its target. It is more efficient than the composite estimator of MSE for about half the districts (52 out of 100), but its performance is more uneven than that of the composite MSE estimator. In principle, the EB-naïve estimator could be improved by combining it with the averaged estimator; however, only minor improvement is made even in the congenial setting (known μ , σ_w^2 and σ_B^2), and the composition is detrimental for several districts in the less congenial settings. Details are omitted.



(in ascending order of sample size)



(in ascending order of sample size)

Figure 4 The mean and root-MSE of estimators of the MSE of the empirical Bayes small-area estimators; estimation of the means of $Y^2/100$

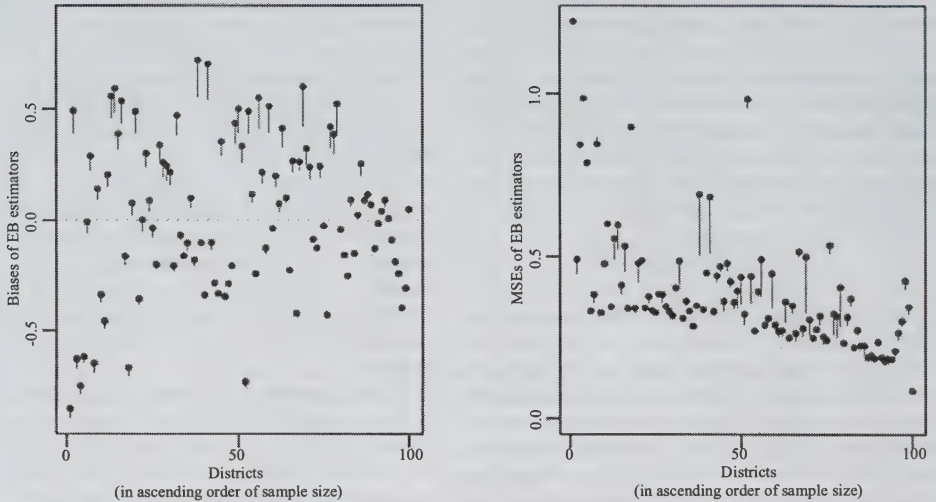


Figure 5 The biases and MSEs of estimators of v_d . The vertical segments connect the quantities associated with \hat{v}_d^* and \hat{v}_d . The quantities associated with \hat{v}_d are marked by black dots

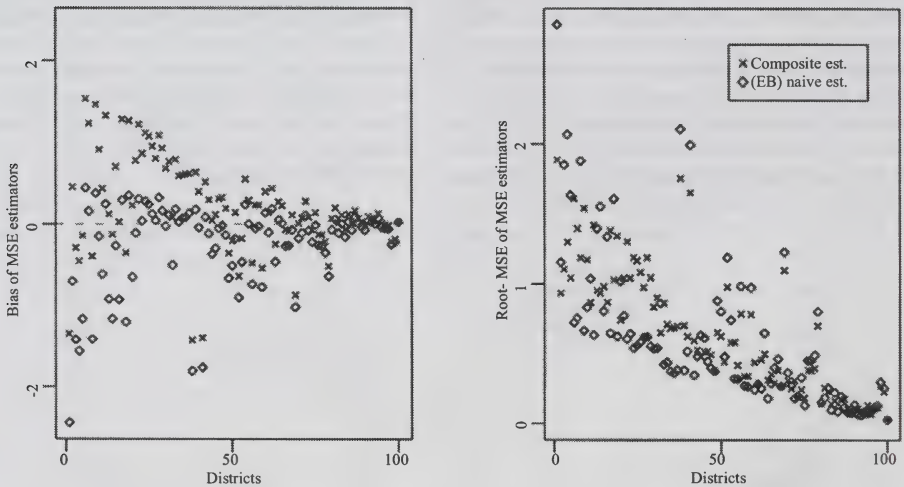


Figure 6 The bias and root-MSE of the composite and empirical-Bayes naive estimators of the MSE of $\bar{\mu}_d$

As a final simulation, we consider a binary outcome variable that indicates whether $Y < 5$, so that the district-level percentages are in the range 1.5–18.8 and the dependence of the percentage on the variance within districts is substantial. The mean of the district-level percentages is 6.85; the substantial skew of these percentages (skewness coefficient equal to 1.01 and kurtosis to 3.78) provides a stern test of the method.

In the simulation, the district-level percentages are estimated by the univariate version of the shrinkage method described in Longford (1999 and 2005, Chapter 8). The results are summarised in Figure 7. The MSE is overestimated by all three estimators for most districts, except for a minority for which the empirical MSE is several times higher than for the rest. The naive estimator has a substantial bias for most districts. The averaged estimator is less

regimented than for normally distributed outcomes because the shrinkage coefficient depends also on the estimated proportion, which is truncated from below at 2% to avoid zero estimated variance $\hat{p}_d(1 - \hat{p}_d)/n_d$. The graph of the composite MSE estimates has the spikes for the appropriate districts, but the spikes are far too short to reduce the bias substantially.

The MSEs of the averaged estimator are satisfactory for most, but are very large for several districts. For the latter districts, the naïve MSE estimator is even less efficient. The composite MSE estimator is less efficient than the averaged estimator for many districts, but the difference is rather small, compensated by the gains in efficiency for districts for which the averaged estimator is less efficient. The EB-naïve MSE estimator resembles in many features the naïve MSE estimator; it is not plotted in the diagram.

In conclusion, this simulation shows that when one of the MSE estimators, in this case the naïve estimator, is very inefficient, it nevertheless contributes, even if very modestly, to the efficiency of the composite MSE estimator. The composite estimator draws on the best that the constituent estimators, averaged and naïve, have to offer, even in uncongenial settings. A remaining challenge is to combine the naïve and averaged estimators to satisfy a particular criterion which trades off the precision for districts that are estimated with high precision for higher precision in estimating in the districts with low precision. For example, we may be less concerned about estimation of the MSEs for districts with abundant representation in the

sample and much more about the sparsely represented districts. Also, some districts (*e.g.*, those in a particular region) may be of specific interest, unrelated to their representation. Of course, the first step in this is the definition of one or a class of criteria that reflect the inferential priorities, and this is bound to be specific to each survey and client. See Longford (2006) for some proposals.

4.1 Refinements and extensions

Several elements of realism can be incorporated in the derivation of the composite MSE estimator. First, uncertainty about μ can be reflected by acknowledging that $\hat{\mu}_d$ and $\hat{\mu}$ are correlated. Thus, $\text{var}(\hat{\mu}_d - \hat{\mu}) = \sigma_w^2(1/n_d - 1/n)$ and the approximation in (5) becomes equality when both instances of σ_w^2/n_d are replaced by $\sigma_w^2(1/n_d - 1/n)$. This brings about only a slight change when $n_d \ll n$, the case for most districts. If the country has a dominant district, with sample size that is a large fraction of the overall sample size, then this adjustment might be relevant, but it has a negligible impact on MSE estimation because even direct estimation of the mean for the district is nearly efficient.

A similar refinement can be applied to the empirical Bayes estimator of μ_d . It amounts to replacing n_d with $1/(n_d^{-1} - n^{-1}) = n_d n / (n - n_d)$ in the coefficient $b_d = 1/(1 + n_d \omega)$. The change is not trivial only for a dominant district, but for such a district shrinkage yields only minute improvement over direct estimation with or without this adjustment.

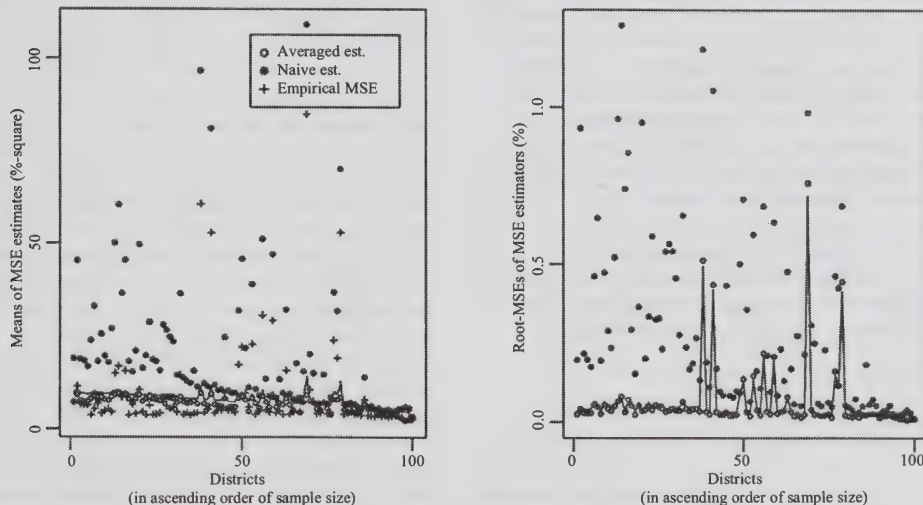


Figure 7 The mean and root-MSE of the composite naïve and averaged estimators of the MSEs of district-level percentages

Accommodating sampling designs that differ from stratified random sampling, and which associate subjects with sampling weights, generates in composite estimation no problems additional to direct estimation with such designs and weights, because we require only the sampling variances of $\hat{\mu}_d$, $\hat{\mu}$ and functions of these. Similarly, exploiting auxiliary information by applying (empirical Bayes) regression

$$y_{jd} = \mathbf{x}_{jd}\boldsymbol{\beta} + \delta_d + \varepsilon_{jd},$$

with independent random samples $\delta_d \sim N(0, \sigma_\delta^2)$ and $\varepsilon_{jd} \sim N(0, \sigma_w^2)$, amounts to replacing $\hat{\mu}$ in (1) with the prediction $\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}$, where $\hat{\mathbf{x}}_d$ is the vector of means of the regressors for district d and $\hat{\boldsymbol{\beta}}$ is the vector of regression parameter estimates. To see this, we express the empirical Bayes fit for district d as

$$\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}} + \frac{n_d\omega}{1+n_d\omega}(\hat{\mu}_d - \hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}) = \frac{n_d\omega}{1+n_d\omega}\hat{\mu}_d + \frac{1}{1+n_d\omega}\hat{\mathbf{x}}_d\hat{\boldsymbol{\beta}}.$$

Pfeffermann *et al.* (1998) discuss issues related to fitting empirical Bayes models to observations with sampling weights. Composite estimation uses direct estimators $\hat{\mu}_d$ and $\hat{\mu}$ for the vectors of all the variables involved and their estimated sampling variance matrices; their evaluation is a standard task in sampling theory. An outstanding problem with empirical Bayes estimators arises when $\hat{\mathbf{x}}_d$ is based on very few observations because the uncertainty about μ_d is then inflated, even when the model fit is very good; if the vector of means \mathbf{x}_d were known (available from sources external to the survey), μ_d would be estimated much more efficiently using $\mathbf{x}_d\boldsymbol{\beta}$. Composite estimation bypasses this problem by searching for the combination of district-level means of auxiliary variables, whether known or estimated from the survey or from other sources, aiming directly to minimise the MSE of the combination (Longford 1999).

The approach developed in Section 3 can be adapted to distributions other than normal straightforwardly, so long as the kurtoses required for evaluating the district-level variance of $(\mu_d - \mu)^2$ and the sampling variance of $(\hat{\mu}_d - \mu)^2$ are known. In practice, kurtosis depends on the mean μ_d , creating difficulties that can be overcome only by approximations or averaging. Estimating proportions p_d with dichotomous data is a case in point. We have

$$\begin{aligned} \text{var}\{(\hat{p}_d - p)^2\} &= \frac{v_d}{n_d^3}(1 - 3p_d + 3p_d^2) \\ &+ \frac{4v_d}{n_d^2}(1 - 2p_d)(p_d - p) + \frac{6v_d}{n_d}(p_d - p)^2 - \frac{v_d^2}{n_d^2}, \end{aligned}$$

where $v_d = p_d(1 - p_d)/n_d$ and p is the national proportion. The complex dependence on the poorly estimated p_d presents an analytical challenge that does not have a universal solution.

Throughout, we assumed that the value of the variance ratio ω is known. In practice, ω is estimated. It is difficult to take account of the uncertainty about ω analytically, but its impact on estimation of μ_d and $\text{MSE}(\hat{\mu}_d; \mu_d)$ can be assessed by sensitivity analysis which repeats the simulations described in Section 4 for a range of plausible values of ω . As one set of simulations takes about one minute of CPU time, this is a manageable computational task. One difficulty in such an assessment is that with an altered assumed value of ω the estimator $\hat{\mu}_d$ is changed, and so the target of the composite MSE estimator is also changed. An alternative informal approach considers the consequences of under- and over-stating the value of ω . In estimating μ_d it is advisable to err on the side of greater ω , giving more weight to the direct estimator $\hat{\mu}_d$ (Longford 2005, Chapter 8). For estimating the MSE of $\hat{\mu}_d$, we may prefer to err on the side of the more stable averaged estimator. That corresponds to increasing the value of the coefficient c_d^* and, as c_d^* is a decreasing function of ω , to reducing the value of ω used for setting c_d^* . Of course, this should be done in moderation, not to discard the contribution of the naïve estimator of MSE altogether.

5. Conclusion

The approach developed in this paper applies the general idea of shrinkage to estimation of MSE of small-area estimators and reduces the impact of averaging, regarded as undesirable when viewed from the design-based perspective, in which the country's districts have fixed population quantities μ_d . We have focussed on improvement in estimation of the MSE for each district separately. In practice, improvement of estimation for some districts is more important than for others. Many surveys are designed for inferences other than small-area estimation, or take small areas into account in planning only peripherally, and so they may yield more than satisfactory estimators for some districts, typically the most populous ones, and less satisfactory for others, often the sparsely populated districts. In such a setting, relatively higher inferential priority should be ascribed to the latter districts. Shrinkage estimators of small-area means and proportions have this property, and the simulations documented in Section 4 indicate that composite estimation of MSE has a similar property, at least in relation to the averaged estimator.

For a given size of the bias in estimating an MSE, we prefer the positive bias, because we regard understating the precision as statistically 'dishonest', whereas overstating it merely fails to present the estimate in the light it deserves - we undersell the results of our analytical effort. With this perspective, the optimal coefficient c_d in (7) should not be

sought by minimising the MSE of the combination, but by a criterion that regards underestimation of MSE as an error more severe than its overestimation by the same amount. Finding a suitable criterion for this, for which optimisation is tractable, is an open problem. The composite MSE estimator derived in Section 3 tends to overestimate the MSE, but this is not by our design.

We have experimented with ML and REML estimation; in the setting used for the simulations, the differences between the two approaches are minute. The advantage of unbiased estimation of the variance σ_B^2 is lost when $\hat{\sigma}_B^2$ is subjected to a non-linear transformation, and efficiency is maintained by transformations only asymptotically. However, small-area estimation is a quintessentially small-sample problem.

The approach presented in this paper illustrates the universality of the general idea of combining alternative estimators. The composite estimator exploits the strengths and reduces the drawbacks of the constituent estimators. Applying it is not detrimental when one of the estimators is far inferior to the other. As a form of averaging is involved even in the composite MSE estimator, it contributes to its robustness by ameliorating departures from the assumptions made in the theoretical development, such as heteroscedasticity and asymmetric (non-normal) within-district distributions.

Incorporating inferential priorities, in effect, redistributing the precision in estimating the MSEs for the small areas, is an open problem. A similar problem, designing surveys for small-area estimation so as to ensure sufficient precision in the model-based perspective (with averaging) is addressed by Longford (2006).

Acknowledgement

Partial support for the work on this manuscript by Grants SEC2003-04476 and SAB2004-0190 from the Spanish Ministry of Education and Science is acknowledged. Insightful and constructive comments of two referees and an Associate Editor are acknowledged.

References

- EURAREA Consortium. (2004). EURAREA Project Final Reference Volume. Enhancing Small-Area Estimation Techniques to Meet European Needs. Office for National Statistics, London. Available from <http://www.statistics.gov.uk/eurarea>.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society, Series A*, 162, 227-245.
- Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. New York: Springer-Verlag.
- Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1988). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Shen, W., and Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B*, 60, 455-471.

Handling survey nonresponse in cluster sampling

Jun Shao¹

Abstract

In surveys under cluster sampling, nonresponse on a variable is often dependent on a cluster level random effect and, hence, is nonignorable. Estimators of the population mean obtained by mean imputation or reweighting under the ignorable nonresponse assumption are then biased. We propose an unbiased estimator of the population mean by imputing or reweighting within each sampled cluster or a group of sampled clusters sharing some common feature. Some simulation results are presented to study the performance of the proposed estimator.

Key Words: Nonignorable nonresponse; Random-effect-based nonresponse; Imputation; Collapsing clusters.

1. Introduction

Nonresponse exists in most survey problems. The probability of having a nonrespondent in a survey item (variable) y typically depends on the unobserved value of y , which creates a great challenge in handling nonrespondents. Commonly used procedures for handling nonresponse (such as reweighting and imputation) are all based on the assumption that nonresponse is ignorable conditional on an auxiliary variable. More precisely,

$$P(y \text{ is a respondent} | y, z) = P(y \text{ is a respondent} | z), \quad (1)$$

where z is an auxiliary variable whose values are observed for all sampled units in the survey. That is, conditional on z , the value of y and its response status are statistically independent. Assumption (1) is referred to as the unconditional response mechanism by Lee, Rancourt and Särndal (1994). Using the terminology in Rubin (1976), nonresponse under (1) is ignorable conditional on z .

There are situations in which it is difficult to find a variable z to satisfy (1). The purpose of this article is to study a method of handling nonresponse when cluster sampling is used, assuming that a variable z satisfying (1) is not available. In cluster sampling, sampling is carried out in two stages; the first stage sampled units are clusters containing units that are sampled in the second stage. Cluster sampling is used because of economic considerations. It is necessary when no reliable list of the second stage units in the population is available (for example, there is no complete list of people but a list of households is available). Under cluster sampling, the variable of interest y may be decomposed as $y = \mu + b + e$, where μ is an unknown overall mean of y , b is a cluster level random effect (all units in the same cluster share the same random effect b), and e is a within-cluster random effect. In many cases, the dependence of the value of y and its response

status is through the unobserved cluster level random effect b :

$$P(y \text{ is a respondent} | y, b) = P(y \text{ is a respondent} | b), \quad (2)$$

i.e., if b were observed, then we would have assumption (1) with $z = b$. For example, suppose that clusters are households and a single person completes survey forms for all sampled persons in a household. It is likely that the response probability depends on the household level variable b , not on the within household variable e .

Assumption (2) was first used by Wu and Carroll (1988) in a health problem where the clusters have a longitudinal (repeated-measure) structure. They called (2) informative censoring (missing) and proposed a method under some parametric assumptions on the probability $P(y \text{ is a respondent} | b)$ and the distribution of y . Later, Little (1995) called this type of missing mechanism the nonignorable random-coefficient-based missing mechanism. Thus, assumption (2) will be referred to as nonignorable random-effect-based response mechanism. Since b is not observed, response mechanism (2) is actually nonignorable.

For survey data, it is difficult to impose any parametric model on the distribution of y . Furthermore, it is also difficult to fit a parametric model for the response mechanism under (2), since b is not observed. After introducing some details on the sampling design and our assumptions, we propose in Section 2 a method for the estimation of the population mean of y under response mechanism (2), without requiring a parametric model for the response mechanism. It is assumed that y follows a random (cluster) effect model, but there is no parametric assumption on the distribution of y . Results from a simulation study are presented in Section 3 for examining the performance of the proposed estimator. Some discussions are given in the last section.

1. Jun Shao, Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

2. Main results

Let S be a sample of clusters of size n from a population P . Within the i^{th} sampled cluster, let S_i be the second stage sample of size $m_i \geq 2$ from a population P_i . For sampled unit $j \in S_i$, a survey weight w_{ij} is constructed (from the specification of the sampling design) so that when there is no nonresponse, $\hat{Y} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$ is an unbiased estimator of the population total Y on any variable y , i.e., $E_s(\hat{Y} - Y) = 0$, where y_{ij} is the y -value of unit j in cluster i $Y = \sum_{i \in P} \sum_{j \in P_i} y_{ij}$, and E_s is the expectation with respect to repeated sampling.

Let y be the variable of interest. We adopt an imputation model approach, i.e., we assume that each y_{ij} in the population is a random variable with

$$y_{ij} = \mu_i + b_i + e_{ij}, \quad (3)$$

where μ_i is an unknown parameter, b_i is an unobserved cluster level random effect with mean 0 and a finite variance, e_{ij} is an unobserved within cluster random effect with mean 0 and a finite variance, and b_i 's and e_{ij} 's are independent. Note that the distribution of y_{ij} may vary with (i, j) .

Let δ_{ij} be the response indicator for y_{ij} ($\delta_{ij} = 1$ if y_{ij} is a respondent and $\delta_{ij} = 0$ if y_{ij} is a nonrespondent). We adopt the approach in Shao and Steel (1999), i.e., δ_{ij} is defined for every unit in the population and nonresponse mechanism is part of the model. Let δ_i be the vector containing δ_{ij} , $j \in S_i$, and y_i be the vector containing y_{ij} , $j \in S_i$. We assume the following nonignorable random-effect-based response mechanism: for every sample,

$$P_m(\delta_i | b_i, y_i) = P_m(\delta_i | b_i), \quad i \in S, \quad (4)$$

where P_m is the probability with respect to the model and $P_m(\xi | \eta)$ denotes the conditional distribution of ξ given η . That is, conditional on b_i, y_i and δ_i are independent. (Unconditionally, they may be dependent.) We assume that the stochastic mechanism with respect to the model is independent of the sampling mechanism so that $E_s E_m(X) = E_m E_s(X)$ as long as X is integrable, where E_m is the expectation with respect to P_m .

Furthermore, we assume that

$$\text{for any } i \in S, \text{ at least one } \delta_{ij} \text{ is } 1. \quad (5)$$

That is, each cluster has at least one respondent. Without this assumption (or some other assumption), the population total Y may not be estimable. More discussion is given in Section 4.

If we assume ignorable nonresponse, i.e., $P_m(\delta_{ij} = 1 | y_{ij}) = P_m(\delta_{ij} = 1)$, then a commonly used procedure is to

impute each nonrespondent by the mean $\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} y_{ij} / \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij}$, which leads to the following estimator of Y :

$$\begin{aligned} \hat{Y}_r &= \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} y_{ij}, \tilde{w}_{ij} \\ &= w_{ij} \left(\frac{\sum_{i \in S} \sum_{j \in S_i} w_{ij}}{\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij}} \right). \end{aligned} \quad (6)$$

Under assumptions (3)-(5),

$$\begin{aligned} E_s E_m(\hat{Y}_r) &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} (\mu_i + b_i + e_{ij}) \right) \\ &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} \mu_i \right) + E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \tilde{w}_{ij} b_i \right), \end{aligned} \quad (7)$$

where the last equality follows from

$$\begin{aligned} E_m(\delta_{ij} \tilde{w}_{ij} e_{ij}) &= E_m[E_m(\delta_{ij} \tilde{w}_{ij} e_{ij} | b_i)] \\ &= E_m[E_m(\delta_{ij} \tilde{w}_{ij} | b_i) E_m(e_{ij} | b_i)] = 0 \end{aligned} \quad (8)$$

under (4). The first term in (7) is equal to

$$E_s E_m \left[\left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \mu_i \right) \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right) / \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \right) \right]$$

which is approximately equal to (when n is large)

$$\begin{aligned} &E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \mu_i \right) E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right) \\ &= \frac{E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} \right)}{E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \mu_i E_m(\delta_{ij}) \right) E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \right)} \\ &= \frac{E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} E_m(\delta_{ij}) \right)}{E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} E_m(\delta_{ij}) \right)}. \end{aligned}$$

Note that

$$E_s E_m(Y) = E_m(Y) = \sum_{i \in P} \sum_{j \in P_i} \mu_i = E_s \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \mu_i \right).$$

Hence, either $\mu_i = \mu$ for all i or $E_m(\delta_{ij})$ does not depend on (i, j) implies that the expectation of the first term in (7) is approximately equal to the expectation of Y . However, $E_m(\delta_{ij} \tilde{w}_{ij} b_i) \neq 0$ in general, because δ_{ij} and b_i

are dependent. Thus, the second term in (7) is not 0 and, hence, \hat{Y}_r defined by (6) is biased under the nonignorable random-effect-based nonresponse. This bias does not go away asymptotically as $n \rightarrow \infty$ and/or $m_i \rightarrow \infty$ for all i .

Recognizing that the problem with \hat{Y}_r is that imputation is done over the entire sample whereas the nonresponse depends on a cluster level random effect, we can find an unbiased estimator by performing imputation within each cluster. This would have been a natural way of imputing if the cluster random effect b_i were observed. If we impute a nonrespondent y_{ij} in cluster i by the cluster mean $\sum_{j \in S_i} \delta_{ij} w_{ij} y_{ij} / \sum_{j \in S_i} \delta_{ij} w_{ij}$, then the resulting estimator is

$$\hat{Y}_c = \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \bar{w}_{ij} y_{ij},$$

with

$$\bar{w}_{ij} = w_{ij} \left(\sum_{j \in S_i} w_{ij} / \sum_{j \in S_i} \delta_{ij} w_{ij} \right). \quad (9)$$

Assumption (5) ensures that \bar{w}_{ij} is well defined. Note that

$$\begin{aligned} E_s E_m(\hat{Y}_c) &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \bar{w}_{ij} \mu_i \right) + E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} \delta_{ij} \bar{w}_{ij} b_i \right) \\ &= E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} \mu_i \right) + E_s E_m \left(\sum_{i \in S} \sum_{j \in S_i} w_{ij} b_i \right) \\ &= E_m(Y), \end{aligned}$$

where the first equality follows from assumption (3) and the fact that, under assumption (4), result (8) still holds with \bar{w}_{ij} replaced by w_{ij} , the second equality follows from the definition of \bar{w}_{ij} and the fact that μ_i and b_i do not depend on j , and the last equality follows from $E_m(b_i) = 0$. Hence, \hat{Y}_c is an unbiased estimator of Y .

Since imputation is done within each cluster, the estimator defined by (9) seems inefficient when some cluster sample sizes m_i are very small. This worry, however, is not necessary in the case where $w_{ij} = w_i$ for all j (e.g., the second stage sampling is with equal probability). When $w_{ij} = w_i$ for all j , imputation leading to \hat{Y}_c in (9) is actually done in a much larger class, a group of clusters sharing something in common. Let $\bar{\delta}_i = m_i^{-1} \sum_{j \in S_i} \delta_{ij}$ be the response rate within cluster i and let

$$G_l = \{i \in S : m_i = m, \bar{\delta}_i = k/m\}, \quad l = (k, m), k \leq m. \quad (10)$$

For each $l = (k, m)$, G_l in (10) is the group of sample clusters having the same $m_i = m$ and $\bar{\delta}_i = k$. If $w_{ij} = w_i$ for all j , then, for $i \in G_l$ with $l = (k, m)$,

$$\begin{aligned} \bar{w}_{ij} &= w_{ij} \left(\sum_{j \in S_i} w_{ij} / \sum_{j \in S_i} \delta_{ij} w_{ij} \right) \\ &= w_i \left(\sum_{j \in S_i} w_i / \sum_{j \in S_i} \delta_{ij} w_i \right) \\ &= w_i / \bar{\delta}_i \\ &= w_i / (k/m) \\ &= w_i \left(\sum_{i \in G_l} m_i w_i \right) / \left(\sum_{i \in G_l} \frac{k}{m} m_i w_i \right) \\ &= w_i \left(\sum_{i \in G_l} m_i w_i \right) / \left(\sum_{i \in G_l} \bar{\delta}_i m_i w_i \right) \\ &= w_i \left(\sum_{i \in G_l} \sum_{j \in S_i} w_i \right) / \left(\sum_{i \in G_l} \sum_{j \in S_i} \delta_{ij} w_{ij} \right) \\ &= w_{ij} \left(\sum_{i \in G_l} \sum_{j \in S_i} w_{ij} \right) / \left(\sum_{i \in G_l} \sum_{j \in S_i} \delta_{ij} w_{ij} \right). \end{aligned}$$

Therefore, imputation leading to \hat{Y}_c in (9) is actually done within each group G_l when $w_{ij} = w_i$ for all j , i.e., a nonrespondent in S_i is imputed by the sample mean of the respondents in G_l , $\sum_{i \in G_l} \sum_{j \in S_i} \delta_{ij} w_{ij} y_{ij} / \sum_{i \in G_l} \sum_{j \in S_i} \delta_{ij} w_{ij}$.

When w_{ij} varies with j for some i 's, some additional conditions are needed in order to combine clusters. A discussion is given in Section 4.

We end this section with a discussion of variance estimation, since most surveys require a variance estimator for each point estimator. A variance formula or its approximation (as $n \rightarrow \infty$) for \hat{Y}_c can be derived, which may require more details on the sampling design. When the first stage sample size n is large, $m_i \leq m$ for all i and a fixed integer m , and n/N is small, where N is the size of P , we can apply the adjusted jackknife method as described in Rao and Shao (1992). More precisely, we can follow the following steps.

1. Create n jackknife replicates, where the i^{th} replicate is obtained by deleting the i^{th} cluster and adjusting the weights to $w_{kj}^{(i)}$, $k \neq i$, $i = 1, \dots, n$, according to the sampling design. For example, if the first stage sampling is a stratified sampling, then $w_{kj}^{(i)} = w_{kj}$ if k and i are not in the same stratum and $w_{kj}^{(i)} = n_h w_{kj} / (n_h - 1)$ if k and i are in the same stratum h , where n_h is the stratum size.
2. Re-impute the nonrespondents in the i^{th} jackknife replicate using the respondents in the i^{th} jackknife replicate, $i = 1, \dots, n$.
3. Compute $\hat{Y}_{c,i}$ the same as \hat{Y}_c but based on the i^{th} re-imputed jackknife replicate, $i = 1, \dots, n$.
4. Compute the jackknife variance estimator for \hat{Y}_c using a standard jackknife formula (e.g., Shao and Tu 1995, Chapter 6). For example, if the first stage sampling is a stratified sampling with H strata, then a jackknife variance estimator is

$$v = \sum_{h=1}^H \frac{n_h - 1}{n_h} \sum_{i \in S_h} \left(\hat{Y}_{c,i} - \frac{1}{n} \sum_{k \in S} \hat{Y}_{c,k} \right)^2,$$

where S_h is the sample from the h^{th} stratum and n_h is the size of S_h .

3. Simulation results

We now present some results from a simulation study to examine the performance of the estimators \hat{Y}_r and \hat{Y}_c .

We create a finite population similar to the elementary school teacher population in Maricopa County, Arizona (Lohr 1999, pages 446-447). The finite population contains 311 clusters (schools). In each cluster, the second stage units are teachers. The cluster size (the number of teachers) varies from 6 to 59 and, hence, the first stage sampling is an unequal probability sampling with probability proportional to cluster size. The first stage sampling is with replacement and the sample size is 31. The second stage sampling is a simple random sampling of size 6 (for any cluster) without replacement.

For each teacher, the variable of interest is the minutes spent per week in school on preparation. The values of y_{ij} for this variable in the simulation are generated according to model (3), where μ_i is the mean minutes spent per week in school on preparation for the i^{th} school, b_i is a random effect of the i^{th} school, and e_{ij} is a random effect of the j^{th} teacher in the i^{th} school. The values of μ_i 's are the sample means in the data set in Lohr (1999, pages 446-447), which vary from 25.52 to 42.18 with a mean of 33.76 and a median of 33.47. The value of b_i is generated according to $b_i = 8.31(X_i - 2)$, where X_i has the gamma distribution with shape parameter 2 and scale parameter 1. The value of e_{ij} is generated from the normal distribution with mean 0 and standard deviation 2.27. The b_i 's and e_{ij} 's are independently generated. The values of $y_{ij} = \mu_i + b_i + e_{ij}$ are generated in each simulation run so that we can evaluate the biases and standard errors of estimators using joint probability under sampling and models (3)-(5).

For sampled units, nonrespondents are generated according to (4) and (5). That is, each sampled cluster has one respondent and the response status of the rest of the sampled units in each cluster are independently determined by $P(y_{ij} \text{ is missing} | b_i) = e^{b_i - 1} / (1 + e^{b_i - 1})$. The mean non-response probability is 33.76%.

For the estimation of the finite population mean, a simulation of 1,000 runs shows that, when \hat{Y}_r is used, the bias, standard error, and root mean squared error are -2.89, 1.32, and 3.17, respectively, and the relative bias $E(\hat{Y}_r - Y)/E(Y)$ is -8.5%; when \hat{Y}_c is used, the bias, standard error, and root mean squared error are 0.12, 1.81, and 1.82, respectively, and the relative bias $E(\hat{Y}_c - Y)/E(Y)$

is 0.3%. This simulation result supports our theory, *i.e.*, \hat{Y}_c is approximately unbiased but \hat{Y}_r is biased. In this case, \hat{Y}_c has a larger standard error than \hat{Y}_r , but \hat{Y}_r has a much larger root mean squared error than \hat{Y}_c due to its large bias.

4. Discussions

Without the assumption that each sampled cluster has at least one respondent, the population total may not be estimable unless some other assumption is added. Under the nonresponse mechanism (4), when all observations in a cluster are nonrespondents, no information in that cluster can be recovered from observed data in other clusters unless some additional assumption is made. For example, one may assume that the population of clusters with no respondent is similar to that of clusters with 1 respondent, in which case one can collapse clusters by distributing the weights of clusters with 0 respondent to the weights of clusters with 1 respondent. Another approach is to assume a model so that we can extrapolate results to clusters with no respondent.

The results in Section 2 are given for mean imputation. Extensions to some other imputation methods are straightforward. For example, if random hot deck imputation is considered, then our result leads to imputation within clusters (or G_i 's). When there is a covariate x whose values are all observed, our result can be extended to regression imputation with model (3) modified to $y_{ij} = \alpha + \beta x_{ij} + b_i + e_{ij}$. For unit nonresponse, our result can also be applied to re-weighting, *i.e.*, adjusting weights within clusters (or G_i 's).

Our method is imputation model based. We assume random-effect model (3) and random-effect-based response mechanism (4). If model (4) does not hold, then $E_m(\delta_{ij} \tilde{w}_{ij} e_{ij}) \neq 0$ and our estimator \hat{Y}_c has a bias with a magnitude depending on the size of $|E_m(\delta_{ij} \tilde{w}_{ij} e_{ij})|$. Similarly, \hat{Y}_c is not valid if model (3) does not hold.

It is shown in Section 2 that the condition $w_{ij} = w_i$ for all j ensures that imputation is done within each G_i that is the group of clusters with the same size and response rate. For two-stage sampling, this condition is satisfied when the last stage sampling is with equal probability (*e.g.*, simple random sampling without replacement). For three-stage sampling, model (3) should be replaced by $y_{ijk} = \mu_{ij} + b_{ij} + e_{ijk}$ and b_i in (4) should be replaced by b_{ij} . The survey weight w_{ijk} satisfies $w_{ijk} = w_{ij}$ as long as the last stage sampling is with equal probability and our result still holds. In two-stage sampling with w_{ij} varying with j , we may perform imputation within a group of clusters that have the same $E_m(y_{ij} | \delta_i)$. For example, suppose that, in addition to (3)-(5), $\mu_i = \mu$, b_i 's are independent and identically distributed (iid), and conditional on b_i , the components of δ_i are iid. Then $E_m(b_i | \delta_i) = E_m(b_i | \tilde{\delta}_i)$ depending only on

the size of the cluster m_i and $\bar{\delta}_i$. Hence we can perform imputation within each G_i defined by (10).

Acknowledgments

This work was partially supported by the NCI Grant CA53786 and NSF Grant DMS-0404535. The author would like to thank Mr. Lei Xu for programming in the simulation study and two referees for their helpful comments.

References

- Lee, H., Rancourt, E., and Särndal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, New York.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Shao, J., and Steel, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
- Wu, M.C., and Carroll, R.J. (1988). Estimation and comparisons of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.

On an optimal controlled nearest proportional to size sampling scheme

Neeraj Tiwari, Arun Kumar Nigam and Ila Pant¹

Abstract

The concept of 'nearest proportional to size sampling designs' originated by Gabler (1987) is used to obtain an optimal controlled sampling design, ensuring zero selection probabilities to non-preferred samples. Variance estimation for the proposed optimal controlled sampling design using the Yates-Grundy form of the Horvitz-Thompson estimator is discussed. The true sampling variance of the proposed procedure is compared with that of the existing optimal controlled and uncontrolled high entropy selection procedures. The utility of the proposed procedure is demonstrated with the help of examples.

Key Words: Controlled sampling; Non-preferred samples; Quadratic programming; High entropy variance.

1. Introduction

In many situations, some samples may be undesirable due to administrative inconvenience, long distance, similarity of units or cost considerations. Such samples are termed non-preferred samples and the technique for avoiding these samples is known as 'controlled selection' or 'controlled sampling'. This technique, originated by Goodman and Kish (1950) has received considerable attention in recent years due to its practical importance.

The technique of controlled sampling is most appropriate when financial or other considerations make it necessary to select a small number of large first stage units, such as hospitals, firms, schools *etc.*, for inclusion in the study. The main purpose of controlled sampling is to increase the probability of selecting a preferred combination beyond that possible with stratified sampling, whilst simultaneously maintaining the initial selection probabilities for each unit of the population, thus preserving the property of a probability sample. This situation generally arises in field surveys where the practical considerations make selection of some units undesirable but it is necessary to follow probability sampling. Controls may be imposed to secure a proper distribution geographically or otherwise and to ensure adequate sample size for some subgroups of the population. Goodman and Kish (1950) considered the reduction of sampling variances of the key estimates as the principal objective of controlled selection, but they also cautioned that this might not always be attained. A real problem emphasizing the need for controls beyond stratification was also discussed by Goodman and Kish (1950, page 354) with the objective of selecting 21 primary sampling units to represent the North Central States. Hess and Srikantan (1966) used the data for the 1961 universe of nonfederal, short-term general medical hospitals in the United States to illustrate the applications of estimation and variance

formulae for controlled selection. Waterton (1983) used the data available from a postal survey of Scottish school leavers carried out in 1977 to describe the advantages of controlled selection and compare the efficiency of controlled selection with multiple proportionate stratified random sampling (meaning the sampling scheme in which instead of one stratifying variable, many variables each of which is associated with the variable of interest y , are used by cross-classifying the population on the basis of these variables) and found the controlled selection to perform favourably.

Three different approaches have been advanced in the recent literature to implement controlled sampling. These are (i) using typical experimental design configurations, (ii) the method of emptying boxes and (iii) using linear programming approaches. While some researchers have used simple random sampling designs to construct the controlled sampling designs, one of the more popular strategies is the use of IPPS (inclusion probability proportional to size) sampling designs in conjunction with the Horvitz-Thompson (1952) estimator. To construct controlled simple random sampling designs, Chakrabarti (1963) and Avadhani and Sukhatme (1973) proposed the use of balanced incomplete block (BIB) designs with parameters $v = N$, $k = n$ and λ , where N is the population size and n is the sample size. Wynn (1977) and Foody and Hedayat (1977) used the BIB designs with repeated blocks for situations where non-trivial BIB designs do not exist. Gupta, Nigam, and Kumar (1982) studied controlled sampling designs with inclusion probabilities proportional to size and used BIB designs in conjunction with the Horvitz-Thompson estimator of the population total $Y (= \sum_{i=1}^N y_i)$, where y_i is the value of the i^{th} unit of the population, U). Nigam, Kumar and Gupta (1984) used some configurations of different types of experimental designs, including BIB designs, to obtain controlled IPPS sampling plans with the

1. Neeraj Tiwari, Ila Pant, Department of Statistics, Kumaon University, S.S.J. Campus, Almora-263601, India. E-mail: kumarn_amo@yahoo.com; Arun Kumar Nigam, Institute of Applied Statistics & Development Studies, Lucknow-226017, India. E-mail: dr_aknigam@yahoo.com.

additional property $c\pi_i\pi_j \leq \pi_{ij} \leq \pi_i\pi_j$ for all $i \neq j = 1, \dots, N$ and some positive constant c such that $0 < c < 1$, where π_i and π_j denote first and second order inclusion probabilities, respectively. Hedayat and Lin (1980) and Hedayat, Lin, and Stufken (1989) used the method of 'emptying boxes' to construct controlled IPPS sampling designs with the additional property $0 < \pi_{ij} \leq \pi_i\pi_j$, $i < j = 1, \dots, N$. Srivastava and Saleh (1985) and Mukhopadhyay and Vijayan (1996) suggested the use of ' t -designs' to replace simple random sampling without replacement (SRSWOR) designs to construct controlled sampling designs.

All the methods of controlled sampling discussed in the previous paragraph may be carried out manually with varying degrees of laboriousness, but none has exploited the advantage of modern computing. Using the simplex method in linear programming, Rao and Nigam (1990, 1992) proposed optimal controlled sampling designs that minimize the probability of selecting the non-preferred samples, while retaining certain properties of an associated uncontrolled plan. Utilizing the approach of Rao and Nigam (1990, 1992), Sitter and Skinner (1994) and Tiwari and Nigam (1998) used the simplex method in linear programming to solve multi-way stratification problems with 'controls beyond stratification'.

In the present article, we use quadratic programming to propose an optimal controlled sampling design which ensures that the probability of selecting non-preferred samples is exactly equal to zero, rather than minimizing it, without sacrificing the efficiency of the Horvitz-Thompson estimator based on an associated uncontrolled IPPS sampling plan. The idea of 'nearest proportional to size sampling designs', introduced by Gabler (1987), is used to construct the proposed design. The Microsoft Excel Solver of the Microsoft Office 2000 package is used to solve the quadratic programming problem. The applicability of the Horvitz-Thompson estimator to the proposed design is discussed. The true sampling variance of the estimate for the proposed design is empirically compared with the variances of the alternative optimal controlled designs of Rao and Nigam (1990, 1992) and uncontrolled high entropy selection procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In Section 3, some examples are considered to demonstrate the utility of the proposed procedure by comparing the probabilities of non-preferred samples and sampling variances of the estimates. Finally in Section 4, the findings of the paper are summarized.

2. The optimal controlled sampling design

In this section, we use the concept of 'nearest proportional to size sampling designs' to propose an optimal

controlled IPPS sampling design that matches the original π_i values, satisfies the sufficient condition $\pi_{ij} \leq \pi_i\pi_j$ for non-negativity of the Yates-Grundy (1953) form of the Horvitz-Thompson (HT) (1952) estimator of the variance and also ensures that the probability of selecting non-preferred samples is exactly equal to zero. Before coming to the proposed plan, we briefly describe the Misdzuno-Sen and Sampford IPPS designs which will be used in the proposed plan for obtaining the initial IPPS design $p(s)$.

2.1 The Midzuno-Sen and Sampford IPPS designs

To introduce the concept of IPPS designs, we assume that a known positive quantity, x_i , is associated with the value of the i^{th} unit of the population and there is reason to believe that the y_i 's are approximately proportional to x_i 's. Here x_i is assumed to be known for all units of the population and y_i is to be collected for sampled units. In IPPS sampling designs, π_i , the probability of including the i^{th} unit in a sample of size n , is np_i , where p_i is the single draw probability of selecting the i^{th} unit in the population (also known as the normal size measure of unit i), given by

$$p_i = \frac{x_i}{\sum_{j=1}^N x_j}, i = 1, 2, \dots, N.$$

We first describe the Midzuno-Sen IPPS scheme and then discuss Sampford's design.

The Midzuno-Sen (MS) (1952, 1953) scheme has a restriction that the probabilities of selecting the i^{th} unit in the population (p_i 's) must satisfy the condition

$$\frac{1}{n} \cdot \frac{n-1}{N-1} \leq p_i \leq \frac{1}{n}, \quad i = 1, 2, \dots, N. \quad (1)$$

If (1) is satisfied for the p_i values of the population under consideration, we apply the MS scheme to get an IPPS plan with the revised probabilities of selection, p_i^* 's, [also known as revised normal size measures] given by

$$p_i^* = np_i \cdot \frac{N-1}{N-n} - \frac{n-1}{N-n}, \quad i = 1, 2, \dots, N. \quad (2)$$

Now, supposing that the s^{th} sample consists of units i_1, i_2, \dots, i_n , the probability of including these units in the s^{th} sample under the MS scheme is given by

$$p(s) = \pi_{i_1, i_2, \dots, i_n} = \frac{1}{\binom{N-1}{n-1}} (p_{i_1}^* + p_{i_2}^* + \dots + p_{i_n}^*). \quad (3)$$

However, due to restriction (1), the MS plan limits the applicability of the method to units that are rather similar in

size. Therefore, when the initial probabilities do not satisfy the condition of the MS plan, we suggest the use of Sampford's (1967) plan to obtain the initial IPPS design $p(s)$.

Using Sampford's scheme, the probability of including n units i_1, i_2, \dots, i_n in the s^{th} sample is given by

$$p(s) = \pi_{i_1, i_2, \dots, i_n} \\ = n K_n \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_n} \left(1 - \sum_{u=1}^n p_{i_u} \right), \quad (4)$$

where $K_n = (\sum_{i=1}^n L_{n-1} / n!)^{-1}$, $\lambda_i = p_i / (1 - p_i)$ for a set $S(m)$ of $m \leq N$ different units, i_1, i_2, \dots, i_m , and L_m is defined as

$$L_0 = 1, L_m = \sum_{S(m)} \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_m} \quad (1 \leq m \leq N).$$

2.2 The proposed plan

Consider a population of N units. Suppose a sample of size n is to be selected from this population. The single draw selection probabilities of these N units of the population (p_i 's) are known. Let S and S_1 denote respectively, the set of all possible samples and the set of non-preferred samples.

Given the selection probabilities for N units of the population, we first obtain an appropriate uncontrolled IPPS design $p(s)$, such as the Midzuno-Sen (1952, 1953) or Sampford (1967) design, as described in Section 2.1. After obtaining the initial IPPS design $p(s)$, the idea behind the proposed plan is to get rid of the non-preferred samples S_1 by confining ourselves to the set $S - S_1$ by introducing a new design $p_0(s)$ which assigns zero probability of selection to each of the non-preferred samples belonging to S_1 , given by

$$p_0(s) = \begin{cases} \frac{p(s)}{1 - \sum_{s \in S_1} p(s)} & \text{for } s \in S - S_1 \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $p(s)$ is the initial uncontrolled IPPS sampling plan.

Consequently, $p_0(s)$ is no longer an IPPS design. So, applying the idea of Gabler (1987), we are interested in the 'nearest proportional to size sampling design' $p_1(s)$ in the sense that $p_1(s)$ minimizes the directed distance D from the sampling design $p_0(s)$ to the sampling design $p_1(s)$, defined as

$$D(p_0, p_1) = E_{p_0} \left[\frac{p_1 - 1}{p_0} \right]^2 = \sum_{s \in S - S_1} \frac{p_1^2(s)}{p_0(s)} - 1 \quad (6)$$

subject to the following constraints:

- (i) $p_1(s) \geq 0$,
 - (ii) $\sum_{s \in S - S_1} p_1(s) = 1$,
 - (iii) $\sum_{s \ni i} p_1(s) = \pi_i$,
 - (iv) $\sum_{s \ni i, j} p_1(s) > 0$ and
 - (v) $\sum_{s \ni i, j} p_1(s) \leq \pi_i \pi_j$.
- (7)

The ordering of the above five constraints is carried out in accordance with their necessity and desirability. Constraints (i) and (ii) are necessary for any probability sampling design. Constraint (iii), which requires that the selection probabilities in the old and new schemes remain unchanged, which ensures that the resultant design will be IPPS. This constraint is a very strong constraint and it affects the convergence properties of the proposed plan to a great extent. Constraint (iv) is highly desirable because it ensures unbiased estimation of the variance. Constraint (v) is desirable as it ensures the sufficient condition for non-negativity of the Yates-Grundy estimator of the variance.

The solution to the above quadratic programming problem, viz., minimizing the objective function (6) subject to the constraints (7), provides us with the optimal controlled IPPS sampling plan that ensures zero probability of selection for the non-preferred samples. The proposed plan is as near as possible to the controlled design $p_0(s)$ defined in (5) and at the same time it achieves the same set of first order inclusion probabilities π_i , as for the original uncontrolled IPPS sampling plan $p(s)$. Due to the constraints (iv) and (v) in (7), the proposed plan also ensures the conditions $\pi_{ij} > 0$ and $\pi_{ij} \leq \pi_i \pi_j$ for the Yates-Grundy estimator of the variance to be stable and non-negative.

The distance measure $D(p_0, p_1)$ defined in (6) is similar to the χ^2 -statistic often employed in related problems and is also used by Cassel and Særdal (1972) and Gabler (1987). Other distance measures are also discussed by Takeuchi, Yanai and Mukherjee (1983). An alternative distance measure for the present discussion may be defined as

$$D(p_0, p_1) = \sum_s \frac{(p_0 - p_1)^2}{(p_0 + p_1)}. \quad (8)$$

When applied on different numerical problems considered by us, we found that the use of (8) gave similar results to (6) in convergence and efficiency and so we will give results using (6) as the distance measure.

While all the other controlled sampling plans discussed by earlier authors attempt to minimize the selection

probabilities of the non-preferred samples, the proposed plan completely excludes the possibility of selecting non-preferred samples by ensuring zero probability for them and at the same time it also ensures the non-negativity of the Yates-Grundy estimator of the variance. However, in some situations a feasible solution to the quadratic programming problem, satisfying all the constraints in (7), may not exist. Constraint (v) may then be relaxed. This may not guarantee the non-negativity of the Yates-Grundy form of the variance estimator. However, since the condition $\pi_{ij} \leq \pi_i \pi_j$ is sufficient for non-negativity of the Yates-Grundy estimator of the variance but not necessary for $n > 2$, as pointed out by Singh (1954), there will still be a possibility of obtaining a non-negative estimator of the variance. After relaxing the constraint (v) in (7), if the Yates-Grundy estimator of the variance comes out to be negative, an alternative variance estimator may be used. This has been demonstrated in Example 5 in Section 3. If even after relaxing constraint (v), a feasible solution of the quadratic programming problem is not found, constraint (iv) may also be relaxed and consequently an alternative variance estimator in place of the Yates-Grundy form of the HT variance estimator may be used. The effect of relaxing these constraints on efficiency of the proposed design is difficult to study, as after relaxing the non-negativity constraint (v) the Yates-Grundy estimator of the variance does not provide accurate results. Using the Yates-Grundy estimator of the variance, for some problems the variance estimate is smaller after relaxing constraint (v) [as in the case of Examples 2(a), 2(b) and 3(a) in Section 3] while for other problems it is larger [as in the case of Example 1(a), 1(b), 3(b), 4(a) and 4(b) in Section 3]. Relaxing a constraint leading to an increased variance estimate may be due to the inability of the Yates-Grundy form of the variance estimator to estimate the true sampling variance correctly, when the non-negativity condition is not satisfied.

The proposed method may also be considered superior to the earlier methods of optimal controlled selection in the sense that setting some samples to have zero selection probability is different from associating a cost with each sample and then trying to minimize the cost, the technique used in earlier approaches of controlled selection. The technique employed by the earlier authors for controlled selection was a crude approach giving some samples very high cost and others very low.

One limitation of the proposed plan is that it becomes impractical when $\binom{N}{n}$ is very large, as the process of enumeration of all possible samples and formation of the objective function and constraints becomes quite tedious. This limitation also holds for the optimum approach of Rao and Nigam (1990, 1992) and other controlled sampling approaches discussed in Section 1. However, with the

advent of faster computing techniques and modern statistical packages, there may not be much difficulty in using the proposed procedure for moderately large populations. On the basis of the size of populations that we have considered in the empirical evaluation, we found that the proposed method can easily handle the controlled selection problems up to a population of 12 units and a sample of size 5. The proposed method may be used to select a small number of first-stage units from each of a large number of strata. This involves a solution of a series of quadratic programming problems, each of a reasonable size, provided the set of non-preferred samples is specified separately in each stratum.

As in the case of linear programming, there is no guarantee of convergence of a quadratic programming problem. Kuhn and Tucker (1951) have derived some necessary conditions for the optimum solution of a quadratic programming algorithm but no sufficient conditions exist for convergence. Therefore unless the Kuhn-Tucker conditions are satisfied in advance, there is no way of verifying whether a quadratic programming algorithm converges to an absolute (global) or relative (local) optimum. Also, there is no way to predict in advance that the solution of a quadratic programming problem exists or not.

2.3 Comparison of sampling variance of the estimate

To estimate the population mean $\bar{Y} (= N^{-1} \sum_{i=1}^N y_i)$ based on a sample s of size n , we use the HT estimator of \bar{Y} defined as

$$\hat{\bar{Y}}_{HT} = \sum_{i \in s} \frac{Y_i}{N\pi_i}. \quad (9)$$

Sen (1953) and Yates and Grundy (1953) showed independently that for fixed size sampling designs, $\hat{\bar{Y}}_{HT}$ has the variance

$$V(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i < j=1}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2, \quad (10)$$

and an unbiased estimator of $V(\hat{\bar{Y}}_{HT})$ is given as

$$\hat{V}(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i < j=1}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2. \quad (11)$$

Constraint (v), when used in the proposed plan, ensures the non-negativity of the variance estimator (11).

To demonstrate the utility of the proposed procedure, we use the empirical examples given in Section 3 to compare the true sampling variance of the HT estimator for the proposed procedure obtained through (10) with variances of the HT estimator using the optimal controlled plan of Rao and Nigam (1990, 1992) and those of two uncontrolled high entropy (meaning the absence of any detectable pattern or

ordering in the selected sample units) procedures of Goodman and Kish (1950) and Brewer and Donadio (2003). In what follows, we reproduce the expressions for the variances of these two high entropy procedures.

The expression for variance of \hat{Y}_{HT} correct to $O(N^{-2})$ using the procedure of Goodman and Kish (1950) is given as

$$V(\hat{Y}_{HT})_{GK} = \frac{1}{nN^2} \left[\sum_{i \in U} p_i A_i^2 - (n-1) \sum_{i \in U} p_i^2 A_i^2 \right] - \frac{n-1}{nN^2} \times \left[2 \sum_{i \in U} p_i^3 A_i^2 - \sum_{i \in U} p_i^2 \sum_{i \in U} p_i^2 A_i^2 - 2 \left(\sum_{i \in U} p_i^2 A_i \right)^2 \right], \quad (12)$$

where $A_i = Y_i / p_i - Y$, $Y = \sum_{i=1}^N Y_i$ and U denotes the finite population of N units.

Recently, Brewer and Donadio (2003) derived the π_{ij} -free formula for the high entropy variance of the HT estimator. They showed that the performance of this variance estimator, under conditions of high entropy, was reasonably good for all populations. Their expression for the variance of the HT estimator is given by

$$V(\hat{Y}_{HT})_{BD} = \frac{1}{N^2} \sum_{i \in U} \pi_i (1 - c_i \pi_i) \left(\frac{Y_i}{\pi_i} - \frac{Y}{n} \right)^2, \quad (13)$$

where $c_i = (n-1) / \{n - (2n-1)(n-1)^{-1} \pi_i + (n-1)^{-1} \sum_{k \in U} \pi_k^2\}$ for all $i \in U$, which appears to perform better than the other values of c_i suggested by them.

3. Examples

In this section, we consider some empirical examples to demonstrate the utility of the proposed procedure and compare it with the existing procedures of optimal controlled sampling. In the present discussion, we begin with the Midzuno-Sen (1952, 1953) IPPS design to demonstrate our procedure, as it is relatively easy to compute the probability of drawing every potential sample under this scheme. However, if the conditions of the Midzuno-Sen scheme are not satisfied, we demonstrate that other IPPS sampling without replacement procedures, such as the Sampford (1967) procedure, may also be used to obtain the initial IPPS design $p(s)$. The true sampling variance of the HT estimator under the proposed plan is also compared with that of the existing procedures of optimal controlled selection and uncontrolled high entropy selection procedures given by (12) and (13).

Example 1: Let us consider a population consisting of six villages, borrowed from Hedayat and Lin (1980). The set S

of all possible samples consists of 20 samples each of size $n=3$. Due to the considerations of travel, organization of fieldwork and cost considerations, Rao and Nigam (1990) identified the following 7 samples as non-preferred samples:

123; 126; 136; 146; 234; 236; 246

(a). The Y_i and p_i values associated with the six villages of the population are:

Y_i : 12 15 17 24 17 19
 p_i : 0.14 0.14 0.15 0.16 0.22 0.19

Since the p_i values satisfy the condition (1), we apply the MS scheme (3) to get an IPPS plan with the revised normal size measures (p_i^* 's) given by (2).

Applying the method discussed in Section 2 and solving the resulting quadratic programming problem with the Microsoft Excel Solver of Microsoft Office 2000 package, we obtain the controlled IPPS plan given in Table 1.

Table 1 Optimal controlled IPPS plan corresponding to Midzuno-Sen (MS) and Sampford's (SAMP) schemes for Example 1

s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]	s	$p_1(s)$ [MS]	$p_1(s)$ [SAMP]
124	0.14	0.09	245	0.03	0.12
125	0.03	0.05	256	0.13	0.14
134	0.00	0.00	345	0.02	0.06
135	0.09	0.03	346	0.20	0.10
145	0.03	0.06	356	0.06	0.06
156	0.13	0.07	456	0.06	0.16
235	0.09	0.05			

This plan matches the original π_i values, satisfies the condition $\pi_{ij} \leq \pi_i \pi_j$ and ensures that the probability of selecting non-preferred samples is exactly equal to zero. Obviously, due to the fulfillment of the condition $\pi_{ij} \leq \pi_i \pi_j$, we can apply the Yates-Grundy form of the HT variance estimator for estimating the variance of the proposed plan.

We have also solved the above example, using plan (3) of Rao and Nigam (1990, page 809) with specified π_{ij} 's taken from the Sampford's plan [to be denoted by RN3] and their plan (4) [to be denoted by RN4]. Using the RN3 plan, the probability of non-preferred samples (ϕ) comes out to be 0.155253 and using the RN4 plan with $c=0.005$, ϕ comes out to be zero, whereas the proposed plan always ensures zero probability to non-preferred samples.

The values of the true sampling variance of the HT estimator $V(\hat{Y}_{HT})$ for the proposed plan, the RN3 plan, the RN4 plan, the Randomized Systematic IPPS sampling plan of Goodman and Kish (1950) [to be denoted by GK] and the uncontrolled high entropy sampling plan of Brewer and Donadio (2003) [to be denoted by BD] are produced in the first row of Table 2. It is clear from Table 2 that the

proposed plan yields almost the same value of variance of the HT estimator as yielded by the RN4 plan. The value of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan is slightly higher than those obtained from the RN3, GK and BD plans. This increase in variance may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 2 Values of the true sampling variance of the HT estimator $[V(\hat{\bar{Y}}_{HT})]$ for the Proposed, RN3, RN4, GK and BD plans

$V(\hat{\bar{Y}}_{HT})$	RN3	RN4	GK	BD	PROPOSED PLAN
Ex 1(a)					
$N = 6, n = 3$	2.93	4.02	3.03	2.92	4.06
Ex 1(b)					
$N = 6, n = 3$	4.76	5.07	4.89	4.15	4.78
Ex 2(a)					
$N = 7, n = 3$	4.48	5.01	4.61	4.45	3.56
Ex 2(b)					
$N = 7, n = 3$	11.97	14.52	12.25	11.44	9.49
Ex 3(a)					
$N = 8, n = 3$	4.85	4.29	4.96	4.86	3.90
Ex 3(b)					
$N = 8, n = 3$	7.29	8.43	7.74	7.37	8.17
Ex 4(a)					
$N = 8, n = 4$	3.19	3.46	3.23	3.15	3.75
Ex 4(b)					
$N = 8, n = 4$	2.41	2.53	2.54	2.38	2.25
Ex 5					
$N = 7, n = 4$	3.08	3.93	3.12	3.07	5.10

(b). Now suppose that the p_i values for the above population of 6 units are as follows:

$p_i:$ 0.10 0.15 0.10 0.20 0.27 0.18

Since these values of p_i do not satisfy the condition (1) of the MS plan, we apply the Sampford (1967) plan to get the initial IPPS design $p(s)$ using (4).

Applying the method discussed in Section 2 and solving the resultant quadratic programming problem, we obtain the controlled IPPS plan given in Table 1. This plan again ensures zero probability to non-preferred samples and satisfies the non-negativity condition for the Yates-Grundy form of the HT variance estimator. This example was also solved by the RN3 and RN4 plans. The value of ϕ for the RN3 plan is 0.064135 and the value of ϕ for the RN4 plan with $c = 0.005$ is zero. The proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are

produced in the second row of Table 2. The proposed plan appears to perform better than the RN4 and GK plans and quite close to other plans considered by us.

Further examples were constructed to analyze the performance of the proposed plan. The populations with Y_i and p_i values and the set of non-preferred samples for each population are summarized in the Appendix. The p_i values for Examples 2(a), 3(a) and 4(a) satisfy the condition (1) of Midzuno-Sen plan and hence for these examples the Midzuno-Sen IPPS plan is used to obtain the initial IPPS design $p(s)$. However, for Examples 2(b), 3(b) and 4(b) the p_i values do not satisfy this condition and therefore we apply the Sampford IPPS plan to obtain the initial IPPS design. The probabilities of non-preferred samples (ϕ) for these examples using the RN3 plan, the RN4 plan and the proposed method are produced in Table 3. Table 3 shows that while the RN3 and RN4 plans only attempt to minimize the probability of non-preferred samples, the proposed plan always ensures zero probability to non-preferred samples.

The values of $V(\hat{\bar{Y}}_{HT})$ for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan for the population summarized in the Appendix are given in Table 2. From Table 2 we conclude that for all the empirical problems considered by us, the proposed plan appears to perform better than or quite close to the RN3, RN4, GK and BD plans. The increase in variance of the estimate for the proposed plan in some cases may be acceptable given the elimination of undesirable samples by the proposed plan.

Table 3 The probabilities of non-preferred samples using RN3, RN4 and Proposed plans

Probability of non-preferred samples (ϕ)	RN3 PLAN	RN4 PLAN	Proposed Plan
Example 2(a)			
$N = 7, n = 3$	0.06	0 ($c = 0.5$)	0
Example 2(b)			
$N = 7, n = 3$	0.05	0 ($c = 0.5$)	0
Example 3(a)			
$N = 8, n = 3$	0.12	0 ($c = 0.005$)	0
Example 3(b)			
$N = 8, n = 3$	0.17	0 ($c = 0.005$)	0
Example 4(a)			
$N = 8, n = 4$	0.05	0 ($c = 0.005$)	0
Example 4(b)			
$N = 8, n = 4$	0.13	0 ($c = 0.005$)	0
Example 5			
$N = 7, n = 4$	0.30	0.1008 ($c = 0.5$)	0

Example 5: We now consider one more example to demonstrate the situation where the proposed plan fails to provide a feasible solution satisfying all the constraints in (7). In such situations, we have to drop a constraint in (7) to obtain a feasible solution of the related quadratic programming problem.

Consider a population of seven villages. Suppose a sample of size $n = 4$ is to be drawn from this population. There are 35 possible samples, out of which the following 14 are considered as non-preferred:

1234;	1236;	1246;	1346;	1357;	1456;	1567;
2345;	2346;	2456;	2567;	3456;	3567;	4567.

Suppose that the following p_i values are associated with the seven villages:

p_i :	0.14	0.13	0.15	0.13	0.16	0.15	0.14.
---------	------	------	------	------	------	------	-------

Since the p_i values satisfy condition (1), we apply the MS plan (3) to obtain the initial IPPS design $p(s)$ and solve the quadratic programming problem by the method discussed in Section 2. However, no feasible solution of the related quadratic programming problem exists in this case. Consequently, we drop constraint (v) in (7) for this particular problem to obtain a feasible solution of the quadratic programming problem. The probabilities of non-preferred samples using the RN3 plan, the RN4 plan and the Proposed plan for this empirical problem are given in the last row of Table 3. The proposed plan again matches the original π_i values and ensures the probability of selecting the non-preferred samples exactly equal to zero. However, due to non-fulfillment of the condition $\pi_{ij} \leq \pi_i \pi_j$ for this example, the non-negativity of the Yates-Grundy estimator of the variance is not ensured. The values of the true variance, $V(\hat{Y}_{HT})$, for the proposed plan, the RN3 plan, the RN4 plan, the GK plan and the BD plan are produced in the last row of Table 2. The value of $V(\hat{Y}_{HT})$ for this empirical example using the proposed plan does not appear to be satisfactory. For such problems where constraint (v) is not satisfied, we suggest the use of alternative variance estimators in place of the Yates-Grundy variance estimator.

We have also solved one more example with $N = 9$ and $n = 4$ using both the Midzuno-Sen and Sampford's methods for obtaining the initial IPPS design $p(s)$. The details of these solutions are omitted for brevity and can be obtained from the authors.

4. Conclusion

We have proposed a quadratic programming approach to solve the controlled sampling problems ensuring zero probability to non-preferred samples. The concept of 'nearest proportional to size sampling designs' of Gabler (1987) is used to obtain the proposed plan. The approach is simple in concept and is very flexible in allowing for a range of different objective functions as well as in permitting a variety of constraints. The only limitation of the procedure is that it cannot be applied to large populations, as the computational process becomes quite tedious for large

populations. The utility of the proposed procedure is demonstrated with the help of examples and its true sampling variance is empirically compared with that of existing controlled sampling plans and uncontrolled high entropy sampling procedures. The proposed plan performs suitably.

Acknowledgements

The authors are grateful to an Associate Editor and two referees for their valuable suggestions and constructive comments on an earlier version of this paper, which led to considerable improvement in presentation of this work.

Appendix

The populations for Example 2-4 with Y_i and p_i values and the set of non-preferred samples.

Example 2. $N = 7, n = 3$.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246; 137; 147; 167; 237; 247; 347; 467.

Y_i :	12	15	17	24	17	19	25
(a). p_i :	0.12	0.12	0.13	0.14	0.20	0.15	0.14
(b). p_i :	0.08	0.08	0.16	0.11	0.24	0.20	0.13

Example 3. $N = 8, n = 3$.

Non-preferred samples: 123; 126; 136; 146; 234; 236; 246; 137; 147; 167; 237; 247; 347; 467; 128; 178; 248; 458; 468; 478; 578.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0.10	0.10	0.11	0.12	0.18	0.13	0.12	0.14
(b). p_i :	0.05	0.09	0.20	0.15	0.10	0.11	0.12	0.18

Example 4. $N = 8, n = 4$.

Non-preferred samples: 1234; 1236; 1238; 1246; 1248; 1268; 1346; 1348; 1357; 1456; 1468; 1567; 1568; 1678; 2345; 2346; 2456; 2468; 2567; 2568; 2678; 3456; 3468; 3567; 3678; 4567; 4678; 5678.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0.11	0.11	0.12	0.13	0.17	0.12	0.11	0.13
(b). p_i :	0.09	0.09	0.18	0.11	0.12	0.14	0.17	0.10

References

- Avadhani, M.S., and Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *International Statistical Review*, 41, 175-182.
- Brewer, K.R.W., and Donadio, M.E. (2003). The high-entropy variance of the Horvitz-Thompson Estimator. *Survey Methodology*, 29, 189-196.
- Cassel, C.M., and Smdal, C.-E. (1972). A model for studying robustness of estimators and informativeness of labels in sampling with varying probabilities. *Journal of Royal Statistical Society, Series B*, 34, 279-289.
- Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *Journal of Indian Statistical Association*, 1, 78-85.
- Foody, W., and Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *Annals of Statistics*, 5, 932-945.
- Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics-Theory & Methods*, 16(4), 1117-1131.
- Goodman, R., and Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of American Statistical Association*, 45, 350-372.
- Gupta, V.K., Nigam, A.K. and Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, 69, 191-196.
- Hedayat, A., and Lin, B.Y. (1980). Controlled probability proportional to size sampling designs. Technical Report, *University of Illinois at Chicago*.
- Hedayat, A., Lin, B.Y. and Stufken, J. (1989). The construction of IPPS sampling designs through a method of emptying boxes. *Annals of Statistics*, 17, 1886-1905.
- Hess, I., and Srikantan, K.S. (1966). Some aspects of probability sampling technique of controlled selection. *Health Serv. Res. Summer* 1966, 8-52.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *Journal of American Statistical Association*, 47, 663-85.
- Kuhn, H.W., and Tucker A.W. (1951). Non-linear programming. *Proceedings of Second Berkely Symposium on Mathematical Statistics and Probability*, 481-492.
- Midzuno, H. (1952). On the sampling system with probability proportional to sums of sizes. *Annals of Institute of Statistics & Mathematics*, 3, 99-107.
- Mukhopadhyay, P., and Vijayan, K. (1996). On controlled sampling designs. *Journal of Statistical Planning & Inference*, 52, 375-378.
- Nigam, A.K., Kumar, P. and Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society, B*, 46, 564-571.
- Rao, J.N.K., and Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.
- Rao, J.N.K., and Nigam, A.K. (1992). 'Optimal' controlled sampling: A unified approach. *International Statistical Review*, 60, 89-98.
- Sampford, M.R. (1967). On sampling with replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.
- Singh, D. (1954). On efficiency of sampling with varying probabilities without replacement. *Journal of Indian Society of Agricultural Statistics*, 6, 48-57.
- Sitter, R.R., and Skinner, C.J. (1994). Multi-way stratification by linear programming. *Survey Methodology*, 20, 65-73.
- Srivastava, J., and Saleh, F. (1985). Need of *t*-designs in sampling theory. *Utilitas Mathematica*, 28, 5-17.
- Takeuchi, K., Yanai, H. and Mukherjee, B.N. (1983). *The Foundations of Multivariate Analysis*. 1st Ed. New Delhi: Wiley Eastern Ltd.
- Tiwari, N., and Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning & Inference*, 69, 89-100.
- Waterton, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.
- Wynn, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics*, 5, 414-418.
- Yates, F., and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society, B*, 15, 253-261.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 22, No. 3, 2006

The Effects of Dependent Interviewing on Responses to Questions on Income Sources Peter Lynn, Anette Jäckle, Stephen P. Jenkins, and Emanuela Sala.....	357
Everyday Concepts and Classification Errors: Judgments of Disability and Residence Roger Tourangeau, Frederick G. Conrad, Zachary Arens, Scott Fricker, Sunghee Lee, and Elisha Smith.....	385
Methods of Behavior Coding of Survey Interviews Yfke P. Ongena and Wil Dijkstra.....	419
Forecasting Labor Force Participation Rates Edward W. Frees	453
Outlier Detection and Editing Procedures for Continuous Multivariate Data Bonnie Ghosh-Dastidar and J.L. Schafer.....	487
A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables Krishnamurty Muralidhar and Rathindra Sarathy.....	507
Record Level Measures of Disclosure Risk for Survey Microdata Elsayed A.H. Elamir and Chris J. Skinner.....	525
Alternative Designs for Regression Estimation Mingue Park.....	541
Variances in Repeated Weighting with and Application to the Dutch Labour Force Survey Paul Kottnerus and Coen van Duin.....	565
The Implication of Employee Stock Options and Holding Gains for Disposable Income and Household Saving Rates in Finland Ilja Kristian Kavonius	585

Volume 22, No. 4, 2006

Ethics, Confidentiality and Data Dissemination Hermann Habermann	599
Discussion Stephen E. Fienberg	615
Discussion Statistics in the National Interest Kenneth Prewitt	621
Discussion Tim Holt	627
Discussion Dennis Trewin	631
Discussion Cynthia Z.F. Clark	637
Discussion Margo Anderson and William Seltzer	641
Rejoinder Hermann Habermann	651
Evaluation of Estimates of Census Duplication Using Administrative Records Information Mary H. Mulry, Susanne L. Bean, D. Mark Bauder, Deborah Wagner, Thomas Mule, and Rita J. Petroni	655
Measuring the Disclosure Protection of Micro Aggregated Business Microdata. An Analysis Taking as an Example the German Structure of Costs Survey Rainer Lenz	681
Statistical Disclosure Control Using Post Randomisation: Variants and Measures for Disclosure Risk Ardo van den Hout and Elsayedh A.H. Elamir	711
A Comparison of Current and Annual Measures of Income in the British Household Panel Survey René Böheim and Stephen P. Jenkins	733
Delete-a-Group Variance Estimation for the General Regression Estimator under Poisson Sampling Phillip S. Kott	759
In Other Journals	769
Editorial Collaborators	773
Index to Volume 22, 2006	777

Volume 34, No. 4, December/décembre 2006

Holger DETTE & Regine SCHEDER Strictly monotone and smooth nonparametric regression for two or more variables	535
Damião N. DA SILVA & Jean D. OPSOMER A kernel smoothing method of adjusting for unit non-response in sample surveys	563
Reinaldo, B. ARELLANO-VALLE & Márcia D. BRANCO & Marc G. GENTON A unified view on skewed distributions arising from selections	581
Stefanie BIEDERMANN, Holger DETTE & Andrey PEPELYSHEV Some robust design strategies for percentile estimation in binary response models	603
Zhide, FANG Some robust designs for polynomial regression models	623
Debbie J. DUPUIS & Maria-Pia VICTORIA-FESER A robust prediction error criterion for Pareto modelling of upper tails	639
Jarrett J. BARBER, Alan E. GELFAND & John A. SILANDER Modelling map positional error to infer true feature location	659
Douglas E. SCHAUBEL & Jianwen CAI Multiple imputation methods for recurrent event data with missing event category	677
José R. BERRENDERO, Antonio CUEVAS & Francisco VÁZQUEZ-GRANDE Testing multivariate uniformity: the distance-to-boundary method	693
Radu HERBEI & Marten H. WEGKAMP Classification with reject option	709
Forthcoming papers/Articles à paraître	730
Online access to The Canadian Journal of Statistics	731
Volume 35 (2007): Subscription rates/Frais d'abonnement	732

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “ $\exp(\cdot)$ ” and “ $\log(\cdot)$ ”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w , ω ; o , O , 0 ; l , 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préféablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1. **Présentation**
 - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8 7/8" par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1 1/2 pouce tout autour.
 - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
 - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
 - 1.4 Les remerciements doivent paraître à la fin du texte.
 - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. **Rédaction**
 - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
 - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(-) et log(-) etc.
 - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
 - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
 - 3.5 Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; l, 1).
 - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
4. **Figures et tableaux**
 - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
 - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. **Bibliographie**
 - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
 - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

Volume 34, No. 4, December/décembre 2006

Holger DETTE & Regine SCHEDER	Strictly monotone and smooth nonparametric regression for two or more variables.....	535
Damião N. DA SILVA & Jean D. OPSOMER	A kernel smoothing method of adjusting for unit non-response in sample surveys.....	563
Reinaldo, B. ARELLANO-VALLE & Marcia D. BRANCO & Marc G. GENTON	A unified view on skewed distributions arising from selections	581
Stefanie BIEDERMANN, Holger DETTE & Andrey PEPELYSHEV	Some robust design strategies for percentile estimation in binary response models.....	603
Zhide, FANG	Some robust designs for polynomial regression models.....	623
Debbie J. DUPUIS & Maria-Pia VICTORIA-FESER	A robust prediction error criterion for Pareto modelling of upper tails.....	639
Jarrett J. BARBER, Alan E. GELFAND & John A. SILANDER	Modelling map positional error to infer true feature location	659
Douglas E. SCHAUBEL & Jianwen CAI	Multiple imputation methods for recurrent event data with missing event category.....	677
José R. BERRENDERO, Antonio CUEVAS & Francisco VÁZQUEZ-GRANDE	Testing multivariate uniformity: the distance-to-boundary method.....	693
Radu HERBEI & Marten H. WEGKAMP	Classification with reject option.....	709
	Forthcoming papers/Articles à paraître.....	730
	Online access to The Canadian Journal of Statistics.....	731
	Volume 35 (2007): Subscription rates/Frais d'abonnement	732

Ethics, Confidentiality and Data Dissemination	599
Hermann Habermann	
Discussion	
Stephen E. Fienberg	615
Discussion	
Statistics in the National Interest	
Kenneth Prewitt	621
Discussion	
Tim Holt	627
Discussion	
Dennis Trewin	631
Discussion	
Cynthia Z.F. Clark	637
Discussion	
Margo Anderson and William Selzer	641
Rejoinder	
Hermann Habermann	651
Evaluation of Estimates of Census Duplication Using Administrative Records Information	
Mary H. Mulry, Susanne L. Bean, D. Mark Bauder, Deborah Wagner, Thomas Mule, and Rita J. Petroni	655
Measuring the Disclosure Protection of Micro Aggregated Business Microdata.	
An Analysis Taking as an Example the German Structure of Costs Survey	
Rainer Lenz	681
Statistical Disclosure Control Using Post Randomisation: Variants and Measures for Disclosure Risk	
Ardo van den Hout and Elsayedh A.H. Elamir	711
A Comparison of Current and Annual Measures of Income in the British Household Panel Survey	
René Böheim and Stephen P. Jenkins	733
Delete-a-Group Variance Estimation for the General Regression Estimator under Poisson Sampling	
Phillip S. Kott	759
In Other Journals	769
Editorial Collaborators	773
Index to Volume 22, 2006	777

JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 22, No. 3, 2006

The Effects of Dependent Interviewing on Responses to Questions on Income Sources	357
Peter Lynn, Anette Jackle, Stephen P. Jenkins, and Emanuela Sala	
Everyday Concepts and Classification Errors: Judgments of Disability and Residence	385
Roger Tourangeau, Frederick G. Conrad, Zachary Arens, Scott Fricker, Sungho Lee, and Elisha Smith	
Methods of Behavior Coding of Survey Interviews	419
Yfke P. Ongena and Wil Dijkstra	
Forecasting Labor Force Participation Rates	453
Edward W. Frees	
Outlier Detection and Editing Procedures for Continuous Multivariate Data	487
Bonnie Ghosh-Dastidar and J.L. Schafer	
A Comparison of Multiple Imputation and Data Perturbation for Masking Numerical Variables	507
Krishnamurthy Muralidhar and Rathindra Sarathy	
Record Level Measures of Disclosure Risk for Survey Microdata	525
Elisayed A.H. Elamir and Chris J. Skinner	
Alternative Designs for Regression Estimation	541
Mingue Park	
Variances in Repeated Weighting with and Application to the Dutch Labour Force Survey	565
Paul Knottnerus and Coen van Duin	
The Implication of Employee Stock Options and Holding Gains for Disposable Income and Household Saving Rates in Finland	585
Ilya Kristian Kavonius	

- Hedayat, A., et Lin, B.Y. (1980). Controlled probability proportional to size sampling designs. Rapport technique, *University of Illinois at Chicago*.
- Hedayat, A., Lin, B.Y., et Sturken, J. (1989). The construction of FPS sampling designs through a method of emptying boxes. *Annals of Statistics*, 17, 1886-1905.
- Hess, I., et Sitikanian, K.S. (1966). Some aspects of probability sampling of controlled selection. *Health Serv. Res. Summer 1966*, 8-52.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from finite universes. *Journal of American Statistical Association*, 47, 663-85.
- Kuhn, H.W., et Tucker A.W. (1951). Non-linear programming. *Statistics and Probability*, 481-492.
- Midzuno, H. (1952). On the sampling system with probability proportional to sums of sizes. *Annals of Institute of Statistics & Mathematics*, 3, 99-107.
- Mukhopadhyay, P., et Vijayan, K. (1996). On controlled sampling designs. *Journal of Statistical Planning & Inference*, 52, 375-378.
- Nigam, A.K., Kumar, P., et Gupta, V.K. (1984). Some methods of inclusion probability proportional to size sampling. *Journal of Royal Statistical Society, B*, 46, 564-571.
- Rao, J.N.K., et Nigam, A.K. (1990). Optimal controlled sampling designs. *Biometrika*, 77, 807-814.
- Rao, J.N.K., et Nigam, A.K. (1992). 'Optimal' controlled sampling: A unified approach. *Revue Internationale de Statistique*, 60, 89-98.
- Samford, M.R. (1967). On sampling with replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities. *Journal of Indian Society of Agricultural Statistics*, 5, 119-127.
- Singh, D. (1954). On efficiency of sampling with varying probabilities without replacement. *Journal of Indian Society of Agricultural Statistics*, 6, 48-57.
- Sitter, R.R., et Skinner, C.J. (1994). Stratification multidimensionnelle par programmation linéaire. *Techniques d'enquête*, 20, 69-78.
- Srivastava, J., et Saleh, F. (1985). Need of *r*-designs in sampling theory. *Utilitas Mathematica*, 28, 5-17.
- Takeuchi, K., Yanai, H., et Mukherjee, B.N. (1983). *The Foundations of Multivariate Analysis*. 1^{re} Ed. New Delhi : Wiley Eastern Ltd.
- Tiwari, N., et Nigam, A.K. (1998). On two-dimensional optimal controlled selection. *Journal of Statistical Planning & Inference*, 69, 89-100.
- Waterson, J.J. (1983). An exercise in controlled selection. *Applied Statistics*, 32, 150-164.
- Wynn, H.P. (1977). Convex sets of finite population plans. *Annals of Statistics*, 5, 414-418.
- Yates, F., et Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of Royal Statistical Society, B*, 15, 253-261.

Midzuno-Sen ainsi que de Sampford pour obtenir le plan PPT initial $p(s)$. Les solutions détaillées de ces problèmes, que nous omettons ici pour être brefs, peuvent être obtenues auprès des auteurs.

4. Conclusion

Nous avons proposé une approche de programmation quadratique pour résoudre les problèmes d'échantillonnage contrôlé en assurant que la probabilité de sélection des échantillons non privilégiés soit nulle. Le concept de « plan d'échantillonnage avec probabilité proportionnelle à la taille le plus proche » de Gabler (1987) est utilisé pour obtenir le plan proposé. Conceptuellement simple et très souple, l'approche permet d'utiliser une gamme de fonctions objectif et diverses contraintes. La seule limite de la méthode tient au fait qu'elle ne peut pas être appliquée à de grandes populations, car le processus de calcul devient assez fastidieux dans ces conditions. L'utilité de la méthode proposée est démontrée à l'aide d'exemples et sa variance d'échantillonnage réelle est comparée empiriquement à celle des plans d'échantillonnage contrôlé et des méthodes d'échantillonnage non contrôlé sous grande entropie existants. Le plan proposé donne des résultats satisfaisants.

Remerciements

Les auteurs remercient un rédacteur adjoint et deux examinateurs de leurs suggestions judicieuses et de leurs commentaires constructifs au sujet d'une version antérieure du présent article qui leur ont permis d'améliorer considérablement la présentation de ces travaux.

Annexe

Populations pour les exemples 2 à 4 avec les valeurs de Y_i et p_i et l'ensemble d'échantillons non privilégiés

Exemple 2. $N = 7, n = 3$.

Echantillons non privilégiés :
123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467.

Y_i :	12	15	17	24	17	19	25
(a). p_i :	0,12	0,12	0,13	0,14	0,20	0,15	0,14
(b). p_i :	0,08	0,08	0,16	0,11	0,24	0,20	0,13

Bibliographie

Exemple 3. $N = 8, n = 3$.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,10	0,10	0,11	0,12	0,18	0,13	0,12	0,14
(b). p_i :	0,05	0,09	0,20	0,15	0,10	0,11	0,12	0,18

Exemple 4. $N = 8, n = 4$.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,11	0,11	0,12	0,13	0,17	0,12	0,11	0,13
(b). p_i :	0,09	0,09	0,18	0,11	0,12	0,14	0,17	0,10

Echantillons non privilégiés :
1234; 1236; 1238; 1246; 1248; 1268; 1346;
1348; 1357; 1456; 1468; 1567; 1568; 1678;
2345; 2346; 2456; 2468; 2567; 2568; 2678;
3456; 3468; 3567; 3678; 4567; 4678; 5678.

Echantillons non privilégiés :
123; 126; 136; 146; 234; 236; 246;
137; 147; 167; 237; 247; 347; 467;
128; 178; 248; 458; 468; 478; 578.

Exemple 3. $N = 8, n = 3$.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,10	0,10	0,11	0,12	0,18	0,13	0,12	0,14
(b). p_i :	0,05	0,09	0,20	0,15	0,10	0,11	0,12	0,18

Exemple 4. $N = 8, n = 4$.

Y_i :	12	15	17	24	17	19	25	18
(a). p_i :	0,11	0,11	0,12	0,13	0,17	0,12	0,11	0,13
(b). p_i :	0,09	0,09	0,18	0,11	0,12	0,14	0,17	0,10

Echantillons non privilégiés :
1234; 1236; 1238; 1246; 1248; 1268; 1346;
1348; 1357; 1456; 1468; 1567; 1568; 1678;
2345; 2346; 2456; 2468; 2567; 2568; 2678;
3456; 3468; 3567; 3678; 4567; 4678; 5678.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *Journal of Indian Statistical Association*, 1, 78-85.

Foody, W., et Hedayat, A. (1977). On theory and applications of BIB designs and repeated blocks. *Annals of Statistics*, 5, 932-945.

Gabler, S. (1987). The nearest proportional to size sampling design. *Communications in Statistics-Theory & Methods*, 16(4), 1117-1131.

Goodman, R., et Kish, L. (1950). Controlled selection-a technique in probability sampling. *Journal of American Statistical Association*, 45, 350-372.

Gupta, V.K., Nigam, A.K. et Kumar, P. (1982). On a family of sampling schemes with inclusion probability proportional to size. *Biometrika*, 69, 191-196.

Vaies-Grundy de l'estimateur de la variance de HT. Nous avons également résolu cet exemple à l'aide des plans RN3 et RN4. La valeur de ϕ est égale à 0,064135 pour la plan RN3 et nulle pour le plan RN4 avec $c = 0,005$. Le plan proposé assure systématiquement que la probabilité de sélection des échantillons non privilégiés soit nulle.

Les valeurs de $V(\hat{Y}_{HT})$ pour le plan proposé, le plan RN3, le plan RN4, le plan GK et le plan BD sont présentées à la deuxième ligne du tableau 2. Le plan proposé semble donner de meilleurs résultats que les plans RN4 et GK, et assez proches de ceux produits par les autres plans étudiés ici.

D'autres exemples ont été construits pour analyser les propriétés du plan proposé. Les populations, avec les valeurs de X_i et p_i et l'ensemble d'échantillons non privilégiés pour chaque population, sont résumées en annexe. Les valeurs de p_i pour les exemples 2(a), 3(a) et 4(a) satisfont la condition (1) du plan de Midzuno-Sen et, donc, satisfont la condition (1) du plan de Midzuno-Sen et, donc, Sen pour obtenir le plan PIPT de Midzuno-pour ces exemples, nous utilisons le plan PIPT de Midzuno-pour obtenir le plan PIPT initial $p(s)$. Toutefois, pour les exemples 2(b), 3(b) et 4(b), les valeurs de p_i ne satisfont pas cette condition et, par conséquent, nous appliquons le plan PIPT de Sampford pour obtenir le plan PIPT initial. Les probabilités des échantillons non privilégiés (ϕ) pour ces exemples en utilisant le plan RN3, le plan RN4 et la méthode proposée sont présentées au tableau 3. Ce dernier montre que, alors que les plans RN3 et RN4 visent uniquement à minimiser la probabilité de sélection des échantillons non privilégiés, le plan proposé assure systématiquement que cette probabilité de sélection soit nulle.

Tableau 3 Probabilités de sélection des échantillons non privilégiés en utilisant les plans RN3, RN4 et proposés

Probabilité des échantillons non privilégiés (ϕ)	PLAN RN3	PLAN RN4	Plan proposé
Exemple 2(a) $N = 7, n = 3$ Exemple 2(b) $N = 7, n = 3$	0,06	0 ($c = 0,5$)	0
Exemple 3(a) $N = 8, n = 3$ Exemple 3(b) $N = 8, n = 3$	0,12	0 ($c = 0,005$)	0
Exemple 4(a) $N = 8, n = 4$ Exemple 4(b) $N = 8, n = 4$	0,05	0 ($c = 0,005$)	0
Exemple 5 $N = 7, n = 4$	0,30	0,1008 ($c = 0,5$)	0

Supposons que les valeurs de p_i qui suivent sont associées aux sept villages :

p_i : 0,14 0,13 0,15 0,16 0,15 0,14 0,14

Puisque les valeurs de p_i satisfont la condition (1), nous appliquons le plan MS (3) pour obtenir le plan PIPT initial $p(s)$ et nous résolvons le problème de programmation quadratique par la méthode exposée à la section 2. Toutefois, aucune solution faisable du problème de programmation quadratique connexe n'existe dans ce cas. Par conséquent, nous laissons tomber la contrainte (v) dans (7) pour ce problème particulier afin d'obtenir une solution faisable. Les probabilités de sélection des échantillons non privilégiés lorsqu'on utilise le plan RN3, le plan RN4 et le plan proposé pour ce problème empirique sont présentées à la dernière ligne du tableau 3. De nouveau, le plan proposé produit les valeurs de p_i originales et assure que la probabilité de sélection des échantillons non privilégiés soit exactement égale à zéro. Cependant, comme la contrainte $\pi_{ij} \leq \pi_i$, n'est pas satisfaisante pour cet exemple, la non-négativité de l'estimateur de la variance de Vaies-Grundy, n'est pas assurée. Les valeurs de la variance réelle, $V(\hat{Y}_{HT})$, pour le plan proposé, le plan RN3, le plan GK et le plan BD sont présentées à la dernière ligne du tableau 2. La valeur de $V(\hat{Y}_{HT})$ pour cet exemple empirique en utilisant le plan proposé ne paraît pas être satisfaisante. Pour ce genre de problème, où la contrainte (v) n'est pas satisfaisante, nous proposons d'utiliser d'autres estimateurs de la variance que celui de Vaies-Grundy.

Nous avons également résolu un dernier exemple en prenant $N = 9$ et $n = 4$ et en utilisant les méthodes de

Exemple 1 : Considérons une population constituée de six villages, emplantée à Hedayat et Lin (1980). L'ensemble S de tous les échantillons possibles comprend 20 échantillons de chacun de taille $n = 3$. Compte tenu des contraintes de déplacement, d'organisations du travail sur le terrain et de coût, Rao et Nigam (1990) ont défini les sept échantillons qui suivent comme étant des échantillons non privilégiés :

- a). Les valeurs de Y_i et p_i associées aux six villages de la population sont :
- | | | | | | |
|---------|------|------|------|------|------|
| 123; | 126; | 136; | 146; | 234; | 246 |
| Y_i : | 12 | 15 | 17 | 24 | 17 |
| p_i : | 0.14 | 0.14 | 0.15 | 0.16 | 0.22 |

Puisque les valeurs de p_i satisfont la condition (1), nous appliquons le plan de MS (3) pour obtenir un plan PPT pour lequel les mesures de taille normale révisées (les valeurs p_i^*) sont données par (2).

En appliquant la méthode décrite à la section 2 et en résolvant les problèmes de programmation quadratique résultants à l'aide du Solveur Microsoft Excel du logiciel Microsoft Office 2000, nous obtenons le plan PPT contrôle

donné au tableau 1.

Tableau 1 Plan PPT contrôle optimal correspondant aux plans de Midzuno-Sen (MS) et de Sampford

s	$p(s)$	[MS]	$p(s)$	[SAMP]	s	$p(s)$	[MS]	$p(s)$	[SAMP]
124	0.14	0.09	0.245	0.03	0.12				
125	0.03	0.05	0.256	0.13	0.14				
134	0.00	0.00	0.345	0.02	0.06				
135	0.09	0.03	0.346	0.20	0.10				
145	0.03	0.06	0.356	0.06	0.06				
156	0.13	0.07	0.456	0.06	0.16				
235	0.09	0.05							

Ce plan reproduit les valeurs originales de π_j , satisfait la condition $\pi_j \leq \pi_j^*$ et assure que les probabilités de sélection des échantillons non privilégiés soient exactement égales à zéro. Évidemment, puisque la condition $\pi_j \leq \pi_j^*$ est satisfait, nous pouvons appliquer la forme de Yates-Grundy de l'estimateur de la variance de HT pour estimer la variance du plan proposé.

Nous avons également résolu l'exemple susmentionné en utilisant le plan (3) de Rao et Nigam (1990, page 809) avec les probabilités π_j spécifiées d'après le plan de Sampford [que nous dénoterons RN3] et d'après leur plan (4) [que nous dénoterons RN4]. Si l'on utilise le plan RN3, la probabilité de sélection des échantillons non privilégiés (ϕ) est égale à 0.155253 et si l'on utilise le plan RN4 avec $c = 0.005$, elle est égale à zéro, alors que le plan proposé assure systématiquement que la probabilité de sélection des échantillons non privilégiés soit nulle.

La valeur de la variance d'échantillonnage réelle de l'estimateur HT $[V(\hat{Y}_{HT})]$ pour le plan proposé, le plan

RN3, le plan RN4, le plan d'échantillonnage PPT randomisé systématique de Goodman et Kish (1950) [que nous dénoterons GK] et le plan d'échantillonnage non contrôle sous grande entropie de Brewer et Donadio (2003) [que nous dénoterons BD] sont présentées à la première ligne du tableau 2. L'examen de ce tableau montre clairement que le plan proposé donne presque la même valeur de la variance de l'estimateur HT que le plan RN4. La valeur de $V(\hat{Y}_{HT})$ pour le plan proposé est légèrement plus élevée que celle obtenue pour les plans RN3, GK et BD. Cette augmentation de la variance pourrait être acceptable, étant donné que le plan proposé élimine les échantillons indésirables.

Tableau 2 Valeurs de la variance d'échantillonnage réelle de l'estimateur HT $[V(\hat{Y}_{HT})]$ pour les plans proposés, RN3, RN4, GK et BD

PLAN PROPOSÉ	RN3	RN4	GK	BD
$V(\hat{Y}_{HT})$	2.93	4.02	3.03	2.92
Ex 1(a)	$N = 6, n = 3$	4.76	5.07	4.89
Ex 1(b)	$N = 6, n = 3$	4.78	4.15	4.78
Ex 2(a)	$N = 7, n = 3$	4.48	5.01	4.61
Ex 2(b)	$N = 7, n = 3$	3.56	4.45	3.56
Ex 3(a)	$N = 7, n = 3$	11.97	14.52	12.25
Ex 3(b)	$N = 8, n = 3$	4.85	4.29	4.86
Ex 4(a)	$N = 8, n = 3$	7.29	8.43	7.74
Ex 4(b)	$N = 8, n = 4$	3.19	3.46	3.23
Ex 5	$N = 7, n = 4$	3.08	3.93	3.12

b). Supposons maintenant que les valeurs p_i pour la population susmentionnée de six unités sont les suivantes :

p_i :	0.10	0.15	0.10	0.20	0.27	0.18
Puisque ces valeurs de p_i ne satisfont pas la condition (1) du plan de MS, nous appliquons le plan de Sampford (1967) pour obtenir le plan PPT initial $p(s)$ en utilisant (4).						
En appliquant la méthode décrite à la section 2 et en résolvant le problème de programmation quadratique résultant, nous obtenons le plan PPT contrôle donné au tableau 1. De nouveau, ce plan assure que la probabilité de sélection des échantillons non privilégiés soit nulle et satisfait la condition de non-négativité de la forme de						

Comme dans le cas de la programmation linéaire, il n'existe aucune garantie de convergence d'un problème de programmation quadratique. Kuhn et Tucker (1951) ont établi certaines conditions nécessaires pour obtenir la solution optimale d'un algorithme de programmation quadratique, mais il n'existe aucune condition suffisante pour la convergence. Par conséquent, à moins que les conditions de Kuhn-Tucker soient satisfaites d'avance, il n'existe aucun moyen de vérifier si un algorithme de programmation quadratique converge vers un optimum absolu (global) ou relatif (local). En outre, il n'existe aucun moyen de prédire si la solution d'un problème de programmation quadratique existe ou non.

2.3 Comparaison de la variance d'échantillonnage de l'estimation

Pour estimer la moyenne de population $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ fondée sur un échantillon s de taille n , nous utilisons l'estimateur HT de \bar{Y} défini comme étant

$$\hat{\bar{Y}}_{HT} = \sum_{i \in s} \frac{y_i}{N \pi_i}. \quad (9)$$

Sen (1953), ainsi que Yates et Grundy (1953) ont montré indépendamment que, pour des plans d'échantillonnage à taille fixe, la variance de $\hat{\bar{Y}}_{HT}$ est donnée par

$$V(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i < j}^N (y_i \pi_j - y_j \pi_i) \left(\frac{\pi_i}{Y} - \frac{\pi_j}{Y} \right), \quad (10)$$

et un estimateur sans biais de $V(\hat{\bar{Y}}_{HT})$ est donné par

$$v(\hat{\bar{Y}}_{HT}) = \frac{1}{N^2} \sum_{i < j}^N \frac{\pi_i \pi_j}{\pi_i \pi_j - \pi_{ij}} \left(\frac{\pi_i}{Y} - \frac{\pi_j}{Y} \right). \quad (11)$$

La contrainte (v), quand elle est utilisée dans le plan proposé, assure la non-négativité de l'estimateur de la variance (11).

Pour démontrer l'utilité de la méthode proposée, nous utilisons des exemples empiriques donnés à la section 3 pour comparer la variance d'échantillonnage réelle de l'estimateur HT pour la méthode proposée obtenue grâce à (10) aux variances de l'estimateur HT lors de l'utilisation du plan contrôlé optimal de Rao et Nigam (1990, 1992) et à celles des deux méthodes non contrôlées sous grande entropie (c'est-à-dire en l'absence de toute régularité décelable ou de tout ordonnancement dans les unités échantillonnées) de Goodman et Kish (1950) et de Brewer et Donadio (2003). Nous reproduisons ci-dessous les expressions des variances pour ces deux méthodes à grande entropie.

3. Exemples

À la présente section, nous examinons certains exemples empiriques en vue de démontrer l'utilité de la méthode proposée et comparons cette dernière aux méthodes existantes d'échantillonnage contrôlé optimal. Nous commençons par discuter du plan PPT de Midzuno-Sen (1952, 1953) pour démontrer notre méthode, car il est relativement facile de calculer la probabilité de tirer chaque échantillon possible sous ce plan. Cependant, si les conditions du plan de Midzuno-Sen ne sont pas satisfaites, nous démontrons que d'autres méthodes d'échantillonnage PPT sans remise, comme celle de Sampford (1967), peuvent être utilisées pour obtenir le plan PPT initial $p(s)$. Nous comparons aussi la variance d'échantillonnage réelle de l'estimateur HT sous le plan proposé à celle obtenue pour les méthodes existantes de sélection contrôlée optimale et de sélection non contrôlée sous grande entropie données par (12) et (13).

$$V(\hat{\bar{Y}}_{HT})^{BD} = \frac{1}{N^2} \sum_{i \in U} \pi_i (1 - c_i \pi_i) \left(\frac{\pi_i}{Y} - \frac{\pi_{i-1}}{Y} \right), \quad (13)$$

où $c_i = (n - 1) / \{ (n - 1) - (2n - 1) - 1 \} \pi_i + (n - 1)^{-1} \sum_{k \in U} \pi_k^2$ pour tout $i \in U$, ce qui semble donner de meilleurs résultats que les autres valeurs de c_i qu'ils ont proposées.

L'estimateur HT est donné par

$$\hat{\bar{Y}}_{HT} = \sum_{i \in U} \frac{y_i}{N \pi_i} \quad (12)$$

Récemment, Brewer et Donadio (2003) ont dérivé la formule ne contenant pas π_{ij} pour la variance sous grande entropie de l'estimateur HT. Ils ont montré que les propriétés de cet estimateur de la variance, sous les conditions de grande entropie, étaient raisonnablement bonnes pour toutes les populations. Leur expression de la variance de l'estimateur HT est donnée par

$$V(\hat{\bar{Y}}_{HT})^{OK} = \frac{1}{nN^2} \left[\sum_{i \in U} p_i^2 A_i^2 - (n-1) \sum_{i \in U} p_i^2 A_i^2 - 2 \sum_{i \in U} p_i^2 A_i^2 - 2 \sum_{i \in U} p_i^2 A_i^2 \right] \quad (12)$$

L'expression de la variance de $\hat{\bar{Y}}_{HT}$ correcte jusqu'à l'ordre $O(N^{-2})$ en utilisant la méthode de Goodman et Kish (1950) est donnée par

Alors que tous les plans d'échantillonnage contrôlé de distance,

discutés par les auteurs antérieurs avaient pour objectif de minimiser les probabilités de sélection des échantillons non privilégiés, celui que nous proposons exclut entièrement la possibilité de sélectionner ces échantillons en garantissant que leurs probabilités de sélection soient nulles tout en assurant la non-négativité de l'estimateur de la variance de Yates-Grundy. Cependant, dans certaines situations, il se pourrait qu'aucune solution faisable du problème de programmation quadratique satisfaisant toutes les contraintes énoncées en (7), n'existe. Le cas échéant, la contrainte (v) peut être relâchée. La non-négativité de la forme de Yates-Grundy de l'estimateur de la variance n'isque alors de ne plus être garantie. Cependant, puisque la condition $\pi_j^* \leq \pi_j$, est suffisante pour que cet estimateur soit non négatif, mais qu'elle n'est pas nécessaire pour $n > 2$, comme l'a souligné Singh (1954), il existera encore une possibilité d'obtenir un estimateur non négatif de la variance. Après relâchement de la contrainte (v) en (7), si l'estimateur de la variance de Yates-Grundy est négatif, un autre estimateur de la variance peut être utilisé. Nous le démontrons à l'aide de l'exemple 5 à la section 3. Si, même après le relâchement de la contrainte (v), une solution faisable du problème de programmation quadratique ne

Lorsqu'elle est appliquée aux divers problèmes numériques que nous considérons, nous constatons que l'équation (8) donne des résultats comparables à ceux que l'on obtient en utilisant la convergence et à l'efficacité, si bien que nous présentons les résultats obtenus en utilisant (6) comme mesure

$$(8) \quad \frac{({}^1d + {}^0d)}{{}_z({}^1d - {}^0d)} \sum^s = ({}^1d \quad {}^0d) D$$

La mesure de distance $D(p_0, p_1)$ définie en (6) est semblable à la statistique χ^2 souvent utilisée dans des problèmes apparentés, et a également été utilisée par Cassel et Sæmdal (1972) et par Takeuchi, Yanai et Mukherjee (1983). Dans le contexte de la présente discussion, nous pourrions aussi définir une autre mesure de distance de la forme

sous les contraintes (7), nous donne le plan d'échantillonnage PPT contrôle optimal assurant que la probabilité de sélection soit nulle pour les échantillons non privilégiés. Le plan proposé est aussi proche que possible du plan contrôle $p_0(s)$ défini en (5) tout en produisant le même ensemble de probabilités d'inclusion de premier ordre π_j^1 que le plan d'échantillonnage PPT non contrôle original $p(s)$. Étant donné les contraintes (iv) et (v) dans (7), le plan proposé garantit également que soient satisfaites les conditions $\pi_j^1 > 0$ et $\pi_j^0 \leq \pi_j^1$, assurant que l'estimateur de Yates-

certaines échantillons et un coût très faible à d'autres. L'une des limites du plan proposé est qu'il devient impossible à appliquer lorsque $\binom{N}{n}$ est très grand, car l'énormité de tous les échantillons possibles et la formation de la fonction objective et des contraintes deviennent assez fastidieuses. Cette limite existe aussi pour l'approche optimale de Rao et Nigam (1990, 1992) et d'autres approches d'échantillonnage contrôlé discutées à la section I. Cependant, grâce aux systèmes informatiques plus rapides et aux logiciels statistiques modernes, l'utilisation de la méthode proposée dans le cas de populations moyennement grandes ne devrait pas être trop difficile. Sur la base des tailles de population que nous avons considérées pour l'évaluation empirique, nous constatons que la méthode proposée permet de traiter facilement les problèmes de sélection contrôlée pour une population allant jusqu'à 12 unités et un échantillon allant jusqu'à 5 unités. La méthode proposée peut être utilisée pour sélectionner un petit nombre d'unités de premier degré dans chacune d'un grand nombre de strates. Cela comprend la résolution d'un série de problèmes de programmation quadratique, ayant chacun une taille raisonnable, à condition que l'ensemble d'échantillons non privilégiés soit spécifié séparément dans chaque strate.

La méthode proposée peut aussi être considérée comme supérieure aux méthodes plus anciennes de sélection contrôlée optimale, car elle consiste à imposer que la probabilité de sélection de certains échantillons soit nulle, au lieu d'associer un coût à chaque échantillon, puis à essayer de minimiser le coût, comme cela a été fait lors des approches antérieures de sélection contrôlée. La technique de sélection contrôlée appliquée par les auteurs antérieurs était une approche grossière consistant à attribuer un coût très élevé à

introuvable, on peut aussi relâcher la contrainte (iv) et, par conséquent, utiliser un autre estimateur de la variance à la place de la forme de Yates-Grundy de l'estimateur HT de la variance. L'effet du relâchement de ces contraintes sur l'efficacité du plan proposé est difficile à étudier, car après le relâchement de la contrainte de non-négativité (v), l'estimateur de la variance de Yates-Grundy ne fournit pas de résultats exacts. Lorsqu'on utilise cet estimateur, pour certains problèmes, l'estimation de la variance est plus fiable après le relâchement de la contrainte (v) [comme dans le cas des exemples 2(a), 2(b) et 3(a) à la section 3], tandis que pour d'autres, elle est plus grande [comme dans le cas des exemples 1(a), 1(b), 3(b), 4(a) et 4(b) à la section 3]. Cette remarque concernant la variance de Yates-Grundy est en accord avec l'observation de la variance pour l'estimateur de la forme de Yates-Grundy de l'estimateur de la variance d'estimer correctement la variance d'échantillonnage réelle, lorsque la condition de non-négativité n'est

Le plan de Midzuno-Sen (MS) (1952, 1953) a pour contrainte que les probabilités de sélection de la $i^{\text{ème}}$ unité (p_i) doivent satisfaire la condition

$$\frac{1}{n} \cdot \frac{N-1}{n-1} \leq p_i \leq \frac{n}{1}, \quad i = 1, 2, \dots, N. \quad (1)$$

Si (1) est satisfaite pour les valeurs de p_i étudiées, nous appliquons le plan de MS pour obtenir un plan PPT dont les probabilités de sélection sont révisées, p_i^* , [également appelées mesures révisées de la taille normale] données par

$$p_i^* = n p_i \cdot \frac{N-1}{n-1} - \frac{n}{N-n}, \quad i = 1, 2, \dots, N. \quad (2)$$

Maintenant, en supposant que le $s^{\text{ème}}$ échantillon est constitué des unités i_1, i_2, \dots, i_n , la probabilité d'inclure ces unités dans le $s^{\text{ème}}$ échantillon sous le plan de MS est donnée par

$$p(s) = \pi_{i_1, i_2, \dots, i_n} = \frac{1}{N-1} (p_{i_1}^* + p_{i_2}^* + \dots + p_{i_n}^*). \quad (3)$$

Toutefois, à cause de la contrainte (1), le plan de MS limite l'applicabilité de la méthode à des unités de taille relativement semblable. Par conséquent, lorsque les probabilités initiales ne satisfont pas la condition du plan de MS, nous proposons d'utiliser le plan de Sampford (1967) pour obtenir le plan PPT initial $p(s)$.

En utilisant le plan de Sampford, la probabilité d'inclure n unités i_1, i_2, \dots, i_n dans le $s^{\text{ème}}$ échantillon est donnée par

$$p(s) = \pi_{i_1, i_2, \dots, i_n} = K^n \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_n} (1 - \sum_{i=1}^n p_i^*), \quad (4)$$

où $K_n = (\sum_{i=1}^n p_i^* / (1 - p_i^*))^{-1}$, $\lambda_i = p_i^* / (1 - p_i^*)$ pour un ensemble $S(m)$ de $m \leq N$ unités différentes, i_1, i_2, \dots, i_m , et L_m est définie comme étant

$$L_0 = 1, L_m = \sum_{i=1}^m \lambda_{i_1} \lambda_{i_2} \dots \lambda_{i_m} \quad (1 \leq m \leq N).$$

2.2 Le plan proposé

Considérons une population de N unités. Supposons que l'on doive sélectionner un échantillon de taille n à partir de cette population. Les probabilités de sélection par tirage unique de ces N unités de la population (valeurs de p_i) sont connues. Soit S et S_1 l'ensemble de tous les échantillons possibles et l'ensemble des échantillons non privilégiés, respectivement.

Sachant les probabilités de sélection pour les N unités de la population, nous obtenons d'abord un plan PPT non contrôlé approprié $p(s)$, tel que le plan de Midzuno-Sen

de paramètres $v = N$, $k = n$ et λ , où N est la taille de la population et n , la taille d'échantillon. Wynn (1977) ainsi que Foody et Hedayat (1977) ont utilisés les plans BIF avec blocs répétés dans des situations où des plans BIF non triviaux n'existent pas. Gupta, Nigam et Kumar (1982) ont étudié des plans d'échantillonnage contrôle avec probabilité d'inclusion proportionnelle à la taille et utilise des plans BIF conjugués à l'estimateur d'Horvitz-Thompson du total de population $Y = \sum_{i=1}^N y_i$, où y_i est la valeur de la $i^{\text{ème}}$ unité de population, Kumar et Gupta (1984) ont utilisé certaines configurations de divers types de plans expérimentaux, y compris des plans BIF, pour obtenir des plans d'échantillonnage PPT contrôle ayant la propriété supplémentaire que $c\pi_i\pi_j \leq \pi_j \leq \pi_i\pi_j$ pour tout $i \neq j = 1, \dots, N$ et une constante positive donnée c telle que $0 < c < 1$, où π_i et π_j représentent les probabilités d'inclusion de premier et de deuxième ordres, respectivement. Hedayat et Lin (1980), ainsi que Hedayat, Lin et Sturken (1989) ont utilisé la méthode du « vidage des boîtes » pour construire des plans d'échantillonnage PPT contrôle ayant la propriété supplémentaire que $0 < \pi_j \leq \pi_i\pi_j, i < j = 1, \dots, N$. Srivastava et Saleh (1985), ainsi que Mukhopadhyay et Vijayan (1996) ont proposé de remplacer l'échantillonnage aléatoire simple sans remise (EASSR) par des « t -designs » pour construire des plans d'échantillonnage contrôle.

Toutes les méthodes d'échantillonnage contrôle mentionnées dans le paragraphe qui précède peuvent être appliquées manuellement avec divers degrés de difficulté, mais aucune n'exploite les avantages de l'informatique moderne. Reconnaissant la méthode du simplex en programmation linéaire, Rao et Nigam (1990, 1992) ont proposé des plans d'échantillonnage contrôle optimal qui réduisent au minimum la probabilité de sélectionner les échantillons non privilégiés, tout en retenant certaines propriétés d'un plan non contrôle comme. En suivant l'approche de Rao et Nigam (1990, 1992), Sitter et Skinner (1994), ainsi que Tiwari et Nigam (1998) ont utilisé la méthode du simplex en programmation linéaire pour résoudre des problèmes de stratification multidimensionnelle avec des « contraintes allant au-delà de la stratification ».

Dans le présent article, nous utilisons la programmation

quadratique pour proposer un plan d'échantillonnage contrôle optimal qui assure que la probabilité de sélectionner les échantillons non privilégiés soit exactement égale à zéro, au lieu de la minimiser, sans sacrifier l'efficacité de l'estimateur d'Horvitz-Thompson fondé sur un plan d'échantillonnage PPT non contrôle comme. Nous utilisons la notion de « plan d'échantillonnage proportionnel à la taille le plus proche » introduite par Gable (1987) pour construire le plan proposé. Nous nous servons du Solveur Microsoft Excel du progiciel Microsoft Office 2000 pour résoudre le problème

2. Le plan d'échantillonnage contrôle optimal

À la présente section, nous nous fondons sur le concept de « plan d'échantillonnage avec probabilité de sélection proportionnelle à la taille le plus proche » pour proposer un plan d'échantillonnage PPT contrôle qui produit des probabilités concordant avec les valeurs originales de π_i , satisfaisant la condition suffisante $\pi_j \leq \pi_i\pi_j$ pour que la forme de Yates-Grundy (1953) de l'estimateur d'Horvitz-Thompson (HT) (1952) de la variance soit non négative et assure en outre que la probabilité de sélection des échantillons non privilégiés soit exactement égale à zéro. Avant de discuter du plan proposé, nous décrivons brièvement les plans PPT de Midzuno-Sen et de Sampford que nous utiliserons dans le plan proposé pour obtenir le plan PPT initial $p(s)$.

2.1 Les plans PPT de Midzuno-Sen et de Sampford

Afin d'introduire le concept des plans PPT, nous supposons qu'une quantité positive connue, x_i , est associée à la valeur de la $i^{\text{ème}}$ unité de la population et qu'il existe une raison de croire que les valeurs de y_i sont approximativement proportionnelles aux x_i . Ici, nous supposons que la valeur de x_i est connue pour toutes les unités de la population et que les valeurs de y_i doivent être recueillies pour toutes les unités échantillonnées. Dans le cas des plans d'échantillonnage PPT, π_i , la probabilité d'inclusion de la $i^{\text{ème}}$ unité dans un échantillon de taille n , est égale à $n\pi_i$, où π_i est la probabilité de sélection en un seul tirage de la $i^{\text{ème}}$ unité de la population (également appelée mesure de taille normale de l'unité i) donnée par

$$p_i = \frac{\sum_{j=1}^J x_j}{x_i}, i = 1, 2, \dots, N.$$

Nous commençons par décrire le plan PPT de Midzuno-Sen, puis nous discutons du plan de Sampford.

Plan d'échantillonnage proportionnel à la taille le plus proche contrôle optimal

Neeraj Tiwari, Arun Kumar Nigam et Ila Pant¹

Résumé

Le concept de « plan d'échantillonnage proportionnel à la taille le plus proche » proposé par Gabler (1987) est utilisé en vue d'obtenir un plan d'échantillonnage contrôle optimal assurant que les probabilités de sélection des échantillons privilégiés soient nulles. L'estimation de la variance pour un plan d'échantillonnage contrôle optimal à l'aide de la méthode Yates-Grundy de l'estimateur d'Horvitz-Thompson est discutée. La variance d'échantillonnage réelle de la méthode proposée est comparée à celle des méthodes existantes de sélection contrôle et non contrôle sous grande entropie. L'utilité de la méthode proposée est démontrée au moyen d'exemples.

Mots clés : Échantillonnage contrôle; échantillons non privilégiés; programmation quadratique; variance sous grande entropie.

1. Introduction

Dans de nombreuses situations, certains échantillons peuvent être indésirables, à cause de complications administratives, de l'éloignement, de la similitude des unités ou de questions de coût. Les échantillons de ce genre sont qualifiés de non privilégiés et la méthode pour les éviter est appelée « sélection contrôle » ou « échantillonnage contrôle ». Cette méthode, proposée pour la première fois par Goodman et Kish (1950), a suscité beaucoup d'intérêt ces dernières années à cause de son importance pratique.

La méthode d'échantillonnage contrôle est la plus appliquée lorsque des considérations financières ou autres obligent à sélectionner un petit nombre de grandes unités primaires d'échantillonnage, comme des hôpitaux, des entreprises ou des écoles, en vue de leur inclusion dans l'étude. L'objectif principal de l'échantillonnage contrôle est d'accroître la probabilité de sélectionner une combinaison privilégiée de sorte qu'elle soit supérieure à celle possible par échantillonnage stratifié, tout en maintenant les probabilités de sélection initiales des unités de la population, donc en préservant la propriété d'échantillon probabiliste. Cette situation se présente généralement dans le cas d'enquêtes sur le terrain, où la sélection de certaines unités est indésirable pour des raisons pratiques, mais un échantillonnage probabiliste est nécessaire. Des contraintes peuvent être imposées afin d'assurer que la répartition géographique ou autre des unités soit appropriée et que la taille de l'échantillon soit adéquate pour certains sous-groupes de la population. Goodman et Kish (1950) ont considéré la réduction de la variance d'échantillonnage des estimations clés comme étant l'objectif principal de la sélection contrôle, mais ont aussi mis en garde contre le fait que cela n'est peut-être pas toujours réalisable. Ils ont aussi discuté d'un

Trois approches distinctes ont été proposées dans la littérature récente en vue de mettre en œuvre l'échantillonnage contrôle, à savoir i) l'utilisation de configurations typiques de plan d'expérience, ii) la méthode du vidage de boîtes et iii) le recours à la programmation linéaire. Bien que certains chercheurs soient partisans des plans d'échantillonnage aléatoire simple pour constituer les plus répandues consistant à conjuguer l'échantillonnage avec probabilité d'inclusion proportionnelle à la taille (PPT) et l'estimateur d'Horvitz-Thompson (1952). Pour construire des plans d'échantillonnage aléatoire simple contrôle, l'une des stratégies les plus répandues consiste à utiliser des stratégies d'échantillonnage en blocs incomplets équilibrés (BIE) ayant pour

problème réel destiné à mettre l'accent sur la nécessité d'utiliser des contraintes au-delà de la stratification (Goodman et Kish 1950, page 354) dans le but de sélectionner 21 unités primaires d'échantillonnage pour représenter les États du Centre-Nord. Hess et Srikantan (1966) ont utilisé les données sur l'univers de 1961 des hôpitaux généraux de soins de courte durée non fédéraux des États-Unis pour illustrer les applications des formules d'estimation et de calcul de la variance au cas de la sélection contrôle. Waterton (1983) a utilisé les données provenant d'une enquête par la poste sur les sortants des écoles écossaises réalisée en 1977 pour décrire les avantages de la sélection contrôle et comparer l'efficacité de cette dernière à celle de l'échantillonnage aléatoire stratifié proportionnel multiple (c'est-à-dire le plan d'échantillonnage dans lequel, au lieu d'une variable de stratification unique, on utilise de nombreuses variables individuellement associées à la variable d'intérêt y par classification croisée de la population en fonction de ces variables) et a constaté que la sélection contrôle donnait des résultats favorables.

1. Neeraj Tiwari, Ila Pant, Département de statistiques, Université Kumon, campus S.S.J., Almorat-263601, Inde. Courriel : kumarn_amo@yahoo.com; Arun Kumar Nigam, Institut de développement, Lucknow-226017, Inde. Courriel : dr_aknigam@yahoo.com.

Les résultats de la section 2 sont présentés pour d'autres méthodes d'imputation sont simples. Par exemple, si l'on considère l'imputation hot deck aléatoire, notre résultat mène à l'imputation dans les grappes (ou les G_j). S'il existe une covariable x dont les valeurs sont toutes observées, notre résultat peut être étendu à l'imputation par la régression en utilisant le modèle (3) modifié pour donner $y_{ij} = \alpha + \beta x_{ij} + b_j + e_{ij}$. Dans le cas de la non-réponse partielle, notre résultat peut également être appliqué à la pondération, c'est-à-dire à l'ajustement des poids dans les grappes (ou les G_j).

Notre méthode est basée sur un modèle d'imputation. Nous utilisons le modèle à effet aléatoire hypothétique (3) et le mécanisme de réponse basé sur les effets aléatoires hypothétiques (4). Si le modèle (4) n'est pas vérifié, alors $E_m(\delta_{ij} w_{ij} e_{ij}) \neq 0$ et notre estimateur \hat{Y}_c a un biais dont la grandeur dépend de la taille de $|E_m(\delta_{ij} w_{ij} e_{ij})|$. De même, \hat{Y}_c n'est pas valide si le modèle (3) ne tient pas.

Nous avons montré à la section 2 qu'en imposant la condition $w_{ij} = w_i$ pour tout j , nous nous assurons que l'imputation est faite dans chaque G_j qui est le groupe de grappes ayant les mêmes taille et taux de réponse. Dans le cas de l'échantillonnage à deux degrés, cette condition est satisfaite quand l'échantillonnage de deuxième degré est réalisé avec probabilités égales (par exemple, l'échantillonnage aléatoire simple sans remise). Dans le cas de l'échantillonnage à trois degrés, le modèle (3) doit être remplacé par $y_{ijk} = \mu_{ij} + b_j + e_{ijk}$ et b_j dans (4), par b_{ij} . L'échantillonnage de dernier degré soit exécuté avec probabilités égales et que notre résultat soit encore vérifié. En cas d'échantillonnage à deux degrés avec w_{ij} variant en fonction de j , nous pouvons procéder à l'imputation dans un groupe de grappes qui ont la même $E_m(V_i | \delta_i)$. Par exemple, supposons qu'en plus des conditions (3) à (5), les

Remerciements

Ce travail a été financé partiellement par la subvention CA53786 du NCI et par la subvention DMS-0404535 de la NSF. L'auteur remercie M. Lei Xu de la programmation de l'étude par simulation et deux examinateurs de leurs commentaires constructifs.

Bibliographie

- Lee, H., Rancourt, E., et Samdal, C.-E. (1994). Experiment with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10, 231-243.
- Little, R.J. (1995). Modeling the dropout mechanism in repeated measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, New York.
- Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Shao, J., et Steel, P. (1999). Variance estimation for imputed survey data with non-negligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
- Wu, M.C., et Carroll, R.J. (1988). Estimation and comparisons of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 44, 175-188.

sont dans la même strate h , où n_h est la taille de la

2. Réimprimer les valeurs pour les non-répondants dans la $i^{\text{ème}}$ réplique jackknife en utilisant les valeurs

des répondants dans la $i^{\text{ème}}$ réplique jackknife,

$i = 1, \dots, m$.

3. Calculer $\hat{Y}_{c,i}$ de la même façon que \hat{Y}_c mais en se

fondant sur la $i^{\text{ème}}$ réplique jackknife réimputée,

$i = 1, \dots, m$.

4. Calculer l'estimateur par le jackknife de la variance

degré est stratifié et comporte H strates, alors l'estimateur par le jackknife de la variance est

$$v = \frac{1}{H} \sum_{h=1}^H m_h \left(\hat{Y}_{c,i} - \frac{1}{m} \sum_{k \in S_h} \hat{Y}_{c,k} \right)^2,$$

où S_h est l'échantillon provenant de la $h^{\text{ème}}$ strate et n_h est la taille de S_h .

3. Résultats des simulations

Nous présentons maintenant les résultats d'une étude par simulation effectuée en vue d'examiner les propriétés des estimateurs \hat{Y}_c et \hat{Y}_c^* .

Nous créons une population finie semblable à la population d'instituteurs et institutrices du comté de Maricopa, en Arizona (Lohr 1999, pages 446-447). La

population finie contient 311 grappes (écoles). Dans chaque grappe, les unités de deuxième degré sont les instituteurs et institutrices. La taille de la grappe (le nombre d'instituteurs et institutrices) varie de 6 à 59, si bien que l'échantillonnage de premier degré est un échantillonnage avec probabilités

inégales proportionnelles à la taille de la grappe. L'échantillonnage de premier degré est fait avec remise et la taille de l'échantillon est 31. L'échantillonnage de deuxième degré est un échantillonnage aléatoire simple sans remise de

taille 6 (pour toute grappe).

Pour chaque instituteur ou institutrice, la variable d'intérêt est le nombre de minutes par semaine consacrées

aux travaux de préparation à l'école. Les valeurs de y_{ij} pour cette variable dans la simulation sont générées conformément au modèle (3), où μ_i est le nombre moyen de minutes consacrées par semaine aux travaux de préparation à l'école pour la $i^{\text{ème}}$ école, b_i est un effet aléatoire de la $i^{\text{ème}}$ école, et e_{ij} est un effet aléatoire du ou de la $j^{\text{ème}}$ instituteur ou institutrice dans la $i^{\text{ème}}$ école. Les valeurs de

μ_i sont les moyennes d'échantillon de l'ensemble de données de Lohr (1999, pages 446-447), qui varient de

25,52 à 42,18 avec une moyenne de 33,76 et une médiane de 33,47. La valeur de b_i est générée conformément à $b_i = 8,31(X_i' - 2)$, où X_i' suit la loi gamma dont le paramètre de forme est 2 et le paramètre d'échelle est 1. La valeur de e_{ij} est générée à partir de la loi normale de moyenne 0 et d'écart-type 2,27. Les b_i et les e_{ij} sont générés indépendamment les uns des autres. Les valeurs de $Y_j^* = \mu_i + b_i + e_{ij}$ sont générées dans chaque exécution de la simulation afin de pouvoir évaluer le biais et les erreurs-types des estimateurs en utilisant la probabilité conjointe sous le plan d'échantillonnage et les modèles (3) à (5). Pour les unités échantillonnées, les non-répondants sont générés conformément à (4) et (5). Autrement dit, chaque grappe échantillonnée contient un répondant et la situation de réponse des autres unités échantillonnées dans chaque grappe est déterminée indépendamment par $P(Y_j^*)$ manquant ($b_i = e^{b_i-1}/(1 + e^{b_i-1})$). La probabilité moyenne de non-réponse est de 33,76 %.

Pour l'estimation de la moyenne de population finie, une simulation comportant 1 000 passages machine montre que, si l'on utilise \hat{Y}_c , le biais, l'erreur-type et la racine de l'erreur quadratique moyenne valent -2,89, 1,32 et 3,17, respectivement, et que le biais relatif $E(\hat{Y}_c) - Y/E(X)$ est de -8,5 %; si l'on utilise \hat{Y}_c^* , le biais, l'erreur-type et la racine de l'erreur quadratique moyenne valent 0,12, 1,81 et 1,82, respectivement et le biais relatif $E(\hat{Y}_c^*) - Y/E(X)$ est de 0,3 %. Ce résultat de simulation corrobore notre théorie, c'est-à-dire que \hat{Y}_c^* est approximativement sans biais, mais que \hat{Y}_c est biaisé. Dans ce cas, l'erreur-type de \hat{Y}_c^* est plus grande que celle de \hat{Y}_c , mais la racine de l'erreur quadratique moyenne de \hat{Y}_c^* est nettement plus grande que celle de \hat{Y}_c à cause de son biais important.

4. Discussion

Si nous n'admettons pas l'hypothèse que chaque grappe échantillonnée contient au moins un répondant, il pourrait être impossible d'estimer le total de population, à moins d'ajouter une autre hypothèse. Sous le mécanisme de non-réponse (4), quand toutes les observations dans une grappe sont des non-réponses, aucune information dans cette grappe ne peut être recouverte d'après les données observées dans d'autres grappes, à moins que soit émise une hypothèse supplémentaire. Par exemple, on pourrait supposer que la population de grappes ne contenant aucun répondant est semblable à celle des grappes contenant un répondant, auquel cas on pourrait regrouper les grappes en répartissant les poids de celles ne contenant aucun répondant de la même façon que les poids de celles contenant un répondant. Une autre approche consiste à utiliser un modèle hypothétique permettant d'extrapoler les résultats aux grappes ne contenant pas de répondant.

grande classe, c'est-à-dire un groupe de grappes ayant une caractéristique en commun. Soit $\delta_i = m_i^{-1} \sum_{j \in S_i} \delta_j$ le taux de réponse dans la grappe i et soit

$$G_l = \{i \in S : m_i = m, \delta_i = k/m\}, \quad l = (k, m), \quad k \leq m. \quad (10)$$

Pour chaque $l = (k, m)$, G_l donne par (10) est le groupe de grappes échantillonnées ayant les mêmes $m_i = m$ et $\delta_i = k$. Si $w_j = w_i$ pour tout j , alors, pour $i \in G_l$ avec $l = (k, m)$,

$$\begin{aligned} w_{ij} &= w_j \left(\sum_{j \in S_i} w_j / \sum_{j \in S_i} \delta_j w_j \right) \\ &= w_i \left(\sum_{j \in S_i} w_i / \sum_{j \in S_i} \delta_j w_i \right) \\ &= w_i \left(\sum_{j \in G_l} m_i w_i / \sum_{j \in G_l} \delta_j m_i w_i \right) \\ &= w_i \left(\sum_{j \in G_l} m_i w_i / \sum_{j \in G_l} \delta_j m_i w_i \right) \\ &= w_i / \delta_i \\ &= w_i / (k/m) \\ &= w_i \left(\sum_{j \in S_i} w_j / \sum_{j \in S_i} \delta_j w_j \right) \\ &= w_i \left(\sum_{j \in S_i} w_i / \sum_{j \in S_i} \delta_j w_i \right) \\ &= w_{ij} \left(\sum_{j \in S_i} w_j / \sum_{j \in S_i} \delta_j w_j \right) \end{aligned}$$

Par conséquent, l'imputation donnant Y_c dans (9) est, en fait, effectuée dans chaque groupe G_l quand $w_j = w_i$ pour tout j , c'est-à-dire que la valeur pour un non-répondant dans S_i est imputée par la moyenne d'échantillon pour les répondants dans G_l , $\sum_{j \in G_l} \delta_j w_j Y_j / \sum_{j \in G_l} \delta_j w_j$. Si w_j varie en fonction de j pour un i donné, certaines conditions supplémentaires sont nécessaires afin de combiner les grappes. Ce point est discuté à la section 4. Nous terminons la présente section par une discussion de l'estimation de la variance, puisque pour la plupart des sondages, un estimateur de variance est requis pour chaque estimateur ponctuel. Nous pouvons dériver une formule de la variance ou son approximation (quand $n \rightarrow \infty$) pour Y_c qui pourrait nécessiter plus de renseignements sur le plan d'échantillonnage. Quand la taille de $m_i \leq m$ pour tout i et un premier degré n est grande, que $m_i \leq m$ pour tout i et un nombre entier fixé m , et que n/N est faible, où N est la taille de P , nous pouvons appliquer la méthode corrigée du jackknife décrite dans Rao et Shao (1992). Plus précisément, nous pouvons procéder aux étapes suivantes.

1. Créer n répliques jackknife, où la i -ième réplique est obtenue en supprimant la i -ième grappe et en rajustant les poids à $w_{(i)}^{(k)}$, $k \neq i$, $i = 1, \dots, n$, exemple, si l'échantillonnage de premier degré est stratifié, alors $w_{(i)}^{(k)} = w_i^{(k)}$ si k et i ne sont pas dans la même strate et $w_{(i)}^{(k)} = n_h w_i^{(k)} / (n_h - 1)$ si k et i

Donc, le fait que $\mu_i = \mu$ pour tout i ou que $E_m(\delta_j)$ ne dépende pas de (i, j) implique que l'espérance du premier terme de (7) est approximativement égale à l'espérance de Y_c . Cependant, en général $E_m(\delta_j w_j b_j) \neq 0$, parce que δ_j et b_j sont dépendants. Donc, le deuxième terme de (7) n'est pas égal à 0 et par conséquent, Y_c défini par (6) est biaisé sous le mécanisme de non-réponse non ignorable basée sur un effet aléatoire. Ce biais ne disparaît pas asymptotiquement quand $n \rightarrow \infty$ et (ou) que $m_i \rightarrow \infty$ pour tout i . Étant donné que le biais de Y_c est dû à ce que l'imputation est effectuée sur l'échantillon complet, alors que la non-réponse dépend d'un effet aléatoire au niveau de la grappe, nous pouvons trouver un estimateur sans biais en exécutant l'imputation dans chaque grappe. Cela serait une méthode d'imputation naturelle si l'effet aléatoire de grappe b_j était observé. Si nous corrigeons une non-réponse Y_j dans la grappe i par imputation de la moyenne de grappe $\sum_{j \in S_i} \delta_j w_j Y_j / \sum_{j \in S_i} \delta_j w_j$, alors l'estimateur résultant est

$$Y_c = \sum_{i \in S} \sum_{j \in S_i} \delta_j w_j Y_j, \quad (9)$$

$$\bar{w}_{ij} = w_j \left(\sum_{j \in S_i} w_j / \sum_{j \in S_i} \delta_j w_j \right).$$

Notons que \bar{w}_{ij} est bien défini. Notons

que

$$E_s E_m(Y_c) = E_s E_m \left(\sum_{i \in P} \sum_{j \in S_i} \delta_j w_j \mu_i \right) + E_s E_m \left(\sum_{i \in P} \sum_{j \in S_i} w_j b_j \right) = E_s E_m(Y_c) + E_s E_m \left(\sum_{i \in P} \sum_{j \in S_i} w_j b_j \right)$$

où la première égalité découle de l'hypothèse (3) et du fait que, sous l'hypothèse (4), le résultat (8) est encore vérifié si l'on remplace w_j par \bar{w}_{ij} , la deuxième égalité découle de la définition de \bar{w}_{ij} et du fait que μ_i et b_j ne dépendent pas de j , et la dernière égalité découle de $E_m(b_j) = 0$. Donc, Y_c est un estimateur sans biais de Y_c .

Puisque l'imputation est faite dans chaque grappe, l'estimateur défini par (9) paraît efficace lorsque certaines tailles d'échantillon de grappes m_i sont très faibles. Cependant, le problème ne se pose pas dans le cas où $w_j = w_i$ pour tout j (par exemple, si les probabilités de sélection sont égales lors de l'échantillonnage de deuxième degré). Quand $w_j = w_i$ pour tout j , l'imputation menant à Y_c dans (9) est, en fait, effectuée dans une beaucoup plus

une méthode d'estimation de la moyenne de population de y sous le mécanisme de réponse (2) qui ne nécessite pas de modèle paramétrique pour le mécanisme de réponse. Nous supposons que y suit un modèle à effet aléatoire (de grappe), mais nous ne formulons aucune hypothèse paramétrique concernant la loi de y . À la section 3, nous exposons les résultats d'une étude par simulation réalisée en vue d'étudier les propriétés de l'estimateur proposé. Enfin, nous présentons une discussion à la dernière section.

2. Principaux résultats

Soit S un échantillon de grappes de taille n tiré d'une population P . Dans la i -ième grappe échantillonnée, soit S_i l'échantillon de deuxième degré de taille $m_i \geq 2$ provenant d'une population P_i . Pour l'unité échantillonnée $j \in S_i$, nous construisons un poids de sondage w_{ij} (d'après la spécification du plan d'échantillonnage) tel que, en

l'absence de non-réponse, $Y = \sum_{j \in S_i} w_{ij} Y_{ij}$ est un estimateur sans biais du total de population Y de toute variable y , c'est-à-dire $E_s(Y - X) = 0$, où Y_{ij} est la valeur de y de l'unité j dans la grappe i , $Y = \sum_{i \in P} \sum_{j \in P_i} Y_{ij}$, et E_s est l'espérance sous échantillonnage répété.

Soit y la variable d'intérêt. Nous adoptons une approche avec modèle d'imputation, autrement dit, nous supposons que chaque y_{ij} dans la population est une variable aléatoire telle que

$$y_{ij} = \mu_i + b_i + e_{ij}, \quad (3)$$

où μ_i est un paramètre inconnu, b_i est un effet aléatoire au niveau de la grappe non observé de moyenne 0 et de variance finie, e_{ij} est un effet aléatoire intragappe non observé de moyenne 0 et de variance finie, et les b_i et les e_{ij} sont indépendants. Notons que la loi de y_{ij} peut varier en fonction de (i, j) .

Soit δ_{ij} l'indicateur de réponse pour y_{ij} ($\delta_{ij} = 1$ si y_{ij} est un répondant et $\delta_{ij} = 0$ si y_{ij} est un non-répondant). Nous adoptons l'approche décrite dans Shao et Steel (1999), c'est-à-dire que δ_{ij} est défini pour chaque unité de la population et que le mécanisme de non-réponse fait partie du modèle. Soit δ_j le vecteur contenant δ_{ij} , $j \in S_i$, et y_j le vecteur contenant y_{ij} , $j \in S_i$. Nous énonçons l'hypothèse que le mécanisme de réponse non ignorable basé sur un effet aléatoire est le suivant : pour chaque échantillon,

$$P^m(\delta_i | b_i, y_i) = P^m(\delta_i | b_i), \quad i \in S, \quad (4)$$

où P^m est la probabilité sous le modèle et $P^m(\xi | \eta)$ dénote la loi conditionnelle de ξ sachant η . Autrement dit, sachant b_i, y_i et δ_j sont indépendants. (Inconditionnellement, ils pourraient être dépendants.) Nous supposons que le mécanisme stochastique sous le modèle est indépendant du

Autrement dit, chaque grappe contient au moins un répondant. Sans cette hypothèse (ou une autre hypothèse), il pourrait être impossible d'estimer le total de population Y . Nous présentons une discussion plus approfondie à la section 4.

Si nous supposons que la non-réponse est ignorable, c'est-à-dire que $P^m(\delta_{ij} = 1 | Y_{ij}) = P^m(\delta_{ij} = 1)$, alors une méthode utilisée fréquemment consiste à imputer la valeur manquante pour chaque non-répondant par la moyenne $\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} Y_{ij'} / \sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'}$, ce qui mène à l'estimateur de Y suivant :

$$Y^r = \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} Y_{ij}, \quad w_{ij} = \sum_{j' \in S_i} \sum_{j'' \in S_i} w_{ij'} / \sum_{j' \in S_i} \sum_{j'' \in S_i} \delta_{ij'} w_{ij'}. \quad (5)$$

pour tout $i \in S$, au moins un δ_{ij} est égal à 1.

Sous les hypothèses (3) à (5),

$$E_s(E_s(Y^r)) = E_s(E_s(Y)) = \sum_{i \in S} \sum_{j \in S_i} \delta_{ij} w_{ij} (\mu_i + b_i + e_{ij}), \quad (6)$$

où la dernière égalité découle de

$$E_m(\delta_{ij} w_{ij} e_{ij}) = E_m[E_m(\delta_{ij} w_{ij} e_{ij} | b_i)] = 0 \quad (7)$$

et la dernière égalité découle de

$$E_m(E_m(Y^r | b_i)) = E_m(E_m(Y | b_i)) = 0 \quad (8)$$

sous (4). Le premier terme de (7) est égal à

$$E_s E_m \left[\left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \mu_i \right) \left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \right) \right] / \left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \right) \quad (9)$$

qui est approximativement égal à (quand n est grand)

$$\frac{E_s E_m \left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \mu_i \right) E_s E_m \left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \right)}{E_s E_m \left(\sum_{j \in S_i} \sum_{j' \in S_i} \delta_{ij'} w_{ij'} \right)^2} = \frac{E_s \left(\sum_{j \in S_i} \sum_{j' \in S_i} w_{ij'} E_s \left(\sum_{j'' \in S_i} \delta_{ij''} w_{ij''} \mu_i \right) \right)}{E_s \left(\sum_{j \in S_i} \sum_{j' \in S_i} w_{ij'} E_s \left(\sum_{j'' \in S_i} \delta_{ij''} w_{ij''} \right) \right)} \quad (10)$$

Traitement de la non-réponse dans les sondages en grappes

Jun Shao¹

Résumé

Dans les sondages en grappes, la non-réponse concernant une variable dépend souvent d'un effet aléatoire au niveau de la grappe et n'est donc pas ignorable. Les estimateurs de la moyenne de population obtenus par imputation ou par pondération sous l'hypothèse de non-réponse ignorable sont alors biaisés. Nous proposons un estimateur sans biais de la moyenne de population obtenu par imputation ou par pondération dans chaque grappe échantillonnée ou dans un groupe de grappes échantillonnées ayant une caractéristique commune. Nous présentons certains résultats obtenus par simulation en vue d'étudier les propriétés de l'estimateur proposé.

Mots clés : Non-réponse non ignorable; non-réponse basée sur un effet aléatoire; imputation; regroupement de grappes.

1. Introduction

La non-réponse existe dans la plupart des problèmes de sondage. La probabilité d'avoir un non-répondant à un item (variable) y dépend habituellement de la valeur inobservée de y , ce qui rend fort difficile le traitement des non-réponses. Les méthodes appliquées habituellement (comme la pondération et l'imputation) sont toutes fondées sur l'hypothèse que la non-réponse est ignorable étant donné une variable auxiliaire. Plus précisément,

$$P(y \text{ est un répondant} | z, y) = P(y \text{ est un répondant} | z), \quad (1)$$

où z est une variable auxiliaire dont les valeurs sont observées pour toutes les unités échantillonnées. Autrement dit, sachant z , la valeur de y et sa situation de réponse sont statistiquement indépendantes. L'hypothèse (1) est appelée mécanisme de réponse non confondu par Lee, Rancourt et Samdal (1994). Selon la terminologie de Rubin (1976), la non-réponse sous (1) est ignorable sachant z .

Dans certaines situations, il est difficile de trouver une variable z qui satisfait (1). L'objectif du présent article est d'étudier une méthode de traitement de la non-réponse dans le cas d'un sondage en grappes, en supposant qu'une variable z satisfaisant (1) n'est pas disponible. Le sondage en grappes comporte un échantillonnage à deux degrés; les unités sélectionnées au premier degré sont des grappes contenant des unités qui, à leur tour, sont échantillonnées au deuxième degré. L'échantillonnage en grappes est utilisé pour des raisons économiques. Il est nécessaire lorsque l'on ne dispose d'aucune liste fiable de population de deuxième degré (par exemple, lorsqu'il n'existe aucune liste complète des personnes, mais que l'on dispose d'une liste d'intérêt y peut être décomposée sous la forme $y = \mu + b + e$, où μ est une moyenne globale inconnue de

y , b est un effet aléatoire au niveau de la grappe (toutes les unités d'une grappe donnée ont en commun le même effet aléatoire b) et e est un effet aléatoire intragappe. Dans de nombreux cas, l'instrument de la corrélation de la valeur de la variable y et de la situation de réponse est l'effet aléatoire non observé au niveau de la grappe b :

$$P(y \text{ est un répondant} | y, b) = P(y \text{ est un répondant} | b), \quad (2)$$

c'est-à-dire que, si b était observé, alors nous obtiendrions l'hypothèse (1) avec $z = b$. Par exemple, supposons que les grappes soient les ménages et que dans chacun d'eux-ci, une seule personne réponde au questionnaire pour tous les membres du ménage échantillonné. Il est vraisemblable que la probabilité des réponses dépende de la variable b au niveau du ménage mais non de la variable e intraménage.

L'hypothèse (2) a été la première utilisée par Wu et Carroll (1988) dans la résolution d'un problème relevant du domaine de la santé où les grappes ont une structure longitudinale (mesure répétée). Ils ont donné à (2) le nom de méthode sous certaines hypothèses paramétriques concernant la probabilité $P(y \text{ est un répondant} | b)$ et la loi de y . Plus tard, Little (1995) a donné à ce genre de mécanisme de création de données manquantes le nom de mécanisme un coefficient aléatoire. Donc, nous donnerons à l'hypothèse (2) le nom de mécanisme de réponse non ignorable basé sur un effet aléatoire. Puisque b n'est pas observé, le mécanisme de réponse (2) est effectivement non ignorable. Dans le cas de données d'enquête, il est difficile de postuler un modèle paramétrique pour la loi de y . Il est, de surcroît, difficile d'ajuster un modèle paramétrique au mécanisme de non-réponse sous (2), car b n'est pas observé. Après la présentation de certains détails sur le plan de sondage et de nos hypothèses, nous proposons à la section 2

ont cette propriété, et les simulations décrites à la section 4 indiquent que l'estimation composite de l'EQM a une propriété semblable, du moins en ce qui a trait à l'estimateur moyen.

Pour une taille donnée du biais dans l'estimation d'une EQM, nous préférons un biais positif, parce que nous considérons la sous-estimation de la précision comme statistiquement « malhonnête », tandis que la surestimation ne présente simplement pas l'estimation sous la lumière qu'elle mérite, autrement dit, nous mettons mal en valeur le résultat de notre effort analytique. Dans cette perspective, le coefficient optimal c_d dans (7) ne devrait pas être recherché par minimisation de l'EQM de la combinaison, mais à l'aide d'un critère qui considère la sous-estimation de l'EQM comme une erreur plus grave que sa surestimation dans la même proportion. Trouver un critère approprié pour lequel l'optimisation est soluble est un problème ouvert. L'estimateur composite de l'EQM dérive à la section 3 a tendance à surestimer l'EQM, mais il ne s'agit pas d'un acte délibéré.

Nous avons expérimenté avec l'estimation du maximum de vraisemblance (ML) et du maximum de vraisemblance restreint (REML), dans les conditions utilisées pour les simulations, les différences entre les deux approches sont très faibles. L'avantage de l'estimation sans biais de la variance σ_b^2 est perdu quand $\hat{\sigma}_b^2$ est soumis à une transformation non linéaire, et l'efficacité n'est maintenue qu'asymptotiquement par les transformations. Cependant, l'estimation pour petits domaines est un problème typique-ment de petit échantillon. L'approche décrite dans le présent article illustre l'universalité de l'idée générale consistant à combiner deux estimateurs et réduit les inconvénients des estimateurs composants. Son application n'est pas préjudiciable lorsque l'un des estimateurs est de loin inférieur à l'autre. Une forme de moyennage intervient même dans l'estimateur composite de l'EQM, de sorte qu'elle contribue à sa robustesse en améliorant les écarts par rapport aux hypothèses faites dans le développement théorique, telles que la présence d'hétéroscédasticité et de lois asymétriques (non normales) dans le district.

Remerciements

Les travaux décrits dans ce manuscrit ont été financés en partie par les subventions SEC2003-04476 et SAB2004-0190 du ministère espagnol de l'Éducation et des Sciences. L'auteur remercie deux examinateurs et un rédacteur associé de leurs commentaires éclairés et constructifs.

Bibliographie

- EURAREA Consortium. (2004). EURAREA Project Final Reference Volume. Enhancing Small-Area Estimation Techniques to Meet European Needs. Office for National Statistics, London. Disponible à <http://www.statistics.gov.uk/eurarea>.
- Ghosh, M., et Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.
- Longford, N.T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society, Séries A*, 162, 227-245.
- Longford, N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician*. New York : Springer-Verlag.
- Longford, N.T. (2006). Calcul de la taille de l'échantillon pour l'estimation pour petits domaines. *Techniques d'enquête*, 32, 97-106.
- Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1988). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Séries B*, 60, 23-40.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Shen, W., et Louis, T.A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Séries B*, 60, 455-471.

problèmes dans l'estimation composite que dans l'estimation directe basée sur ce genre de plans et de pondérations, parce que l'estimation composite requiert uniquement les variances d'échantillonnage de $\hat{\mu}_d$, $\hat{\mu}_1$ et de leurs fonctions. De même, l'exploitation de l'information auxiliaire par recours à la régression (empirique bayésienne)

$$y^{jd} = x^{jd} + \delta_d + \varepsilon^{jd}$$

avec des échantillons aléatoires indépendants, $\delta_d \sim N(0, \sigma_{\delta_d}^2)$ et $\varepsilon^{jd} \sim N(0, \sigma_{\varepsilon^{jd}}^2)$, équivalant à remplacer $\hat{\mu}_1$ dans (1) par la prédiction \hat{x}_d^{jd} , où \hat{x}_d^{jd} est le vecteur des moyennes pour le district d et $\hat{\beta}$ est le vecteur des estimations des paramètres de régression. Pour le voir, nous exprimons l'ajustement empirique bayésien pour le district d sous la forme

$$\hat{x}_d^{jd} \hat{\beta} + \frac{1 + n_d \omega}{n_d \omega} (\hat{\mu}_d - \hat{x}_d^{jd} \hat{\beta}) = \frac{1 + n_d \omega}{n_d \omega} \hat{\mu}_d + \frac{1 + n_d \omega}{1} \hat{x}_d^{jd} \hat{\beta}.$$

Prefermann et coll. (1998) discutent des problèmes associés à l'ajustement des modèles empiriques bayésiens aux observations avec poids d'échantillonnage. L'estimation composite s'appuie sur des estimateurs directs $\hat{\mu}_d$ et $\hat{\mu}_1$ pour les vecteurs de toutes les variables concernées et leurs matrices de variances d'échantillonnage estimées; leur évaluation est une tâche standard en théorie de l'échantillonnage. Un problème demeure non résolu dans le cas des estimateurs empiriques bayésiens quand \hat{x}_d^{jd} est basé sur un très petit nombre d'observations, parce que l'incertitude au sujet de $\hat{\mu}_d$ est alors grossie, même si l'ajustement du modèle est très bon; si le vecteur des moyennes \hat{x}_d^{jd} était connu (d'après des sources externes à l'enquête), $\hat{\mu}_d$ pourrait être estimé de manière beaucoup plus efficace en utilisant $\hat{x}_d^{jd} \hat{\beta}$. L'estimation composite permet de contourner ce problème en recherchant la combinaison de moyennes de variables auxiliaires, qu'elles soient connues ou estimées d'après l'enquête ou d'autres sources, visant directement à minimiser l'EQM de la combinaison (Longford 1999).

L'approche élaborée à la section 3 peut être adaptée facilement à d'autres distributions que la loi normale, à condition que les ajustements nécessaires pour évaluer la variance au niveau du district de $(\hat{\mu}_d - \hat{\mu}_1)^2$ et la variance d'échantillonnage de $(\hat{\mu}_d - \hat{\mu}_1)^2$ soient connues. En pratique, l'ajustement dépend de la moyenne $\hat{\mu}_d$, ce qui crée des difficultés qui ne peuvent être surmontées que par des approximations ou des calculs de valeur moyenne. L'estimation de proportions p_d d'après des données dichotomiques est un exemple typique. Nous avons

$$\text{var}\{(\hat{p}_d - p_d)^2\} = \frac{p_d}{V_d} (1 - p_d)^2 + 3p_d^2 \left(\frac{p_d}{V_d} + \frac{p_d}{6V_d} (p_d - p_d)^2 - \frac{p_d^2}{V_d^2} \right) + \frac{p_d^2}{V_d^2}.$$

5. Conclusion

L'approche élaborée dans le présent article applique la notion générale de rétrécissement à l'estimation de l'EQM des estimateurs pour petits domaines et réduit l'effet du moyennage, considéré comme indésirable sous l'angle de l'approche fondée sur le plan de sondage, dans laquelle les quantités de population μ_d des districts du pays sont fixes. Nous nous sommes concentrés sur l'ajustement individuel de l'estimation de l'EQM pour chaque district. En pratique, l'amélioration de l'estimation est plus importante pour certains districts que pour d'autres. De nombreuses enquêtes sont conçues pour faire d'autres inférences que l'estimation pour petits domaines ou ne tiennent compte que périphérieusement des petits domaines dans la planification, mais moins satisfaisants pour d'autres, souvent ceux qui sont peu peuplés. Dans de telles conditions, on devrait accorder une priorité inférentielle relativement plus grande à ces derniers districts. Les estimateurs de rétrécissement des moyennes et des proportions de petits domaines universelle.

Tout au long de l'exposé, nous avons supposé que la valeur du ratio des variances ω est connue. En pratique, ω est estimée. Il est difficile de tenir compte de l'incertitude au sujet de ω analytiquement, mais son effet sur l'estimation de $\hat{\mu}_d$ et de $\text{EQM}(\hat{\mu}_d; \hat{\mu}_d)$ peut être évalué par analyse de sensibilité en répétant les simulations décrites à la section 4 pour une gamme de valeurs plausibles de ω . Puisqu'un ensemble de simulations requiert environ une minute de temps de processeur, il s'agit d'une tâche de calcul faisable. Une difficulté de ce genre d'évaluation tient au fait qu'avec une valeur hypothétique altérée de ω , l'estimateur $\hat{\mu}_d$ est changé, et, donc, la cible de l'estimateur composite de l'EQM est changée également. Une approche de rechange informelle consiste à considérer les conséquences d'une sous-estimation et d'une surestimation de la valeur de ω . Lors de l'estimation de $\hat{\mu}_d$, il est conseillé d'erreur dans la direction d'une valeur de ω plus grande, dominant plus de poids à l'estimateur direct $\hat{\mu}_d$ (Longford 2005, chapitre 8). Pour estimer l'EQM de $\hat{\mu}_d$, nous pourrions préférer erreur dans la direction de l'estimateur moyen, plus stable. Cela revient à accroître la valeur du coefficient c_d^* et, comme c_d^* est une fonction décroissante de ω , à réduire la valeur de ω utilisée pour fixer c_d^* . Naturellement, la modération est de rigueur, afin de ne pas écartier entièrement la contribution de l'estimateur naïf de l'EQM.

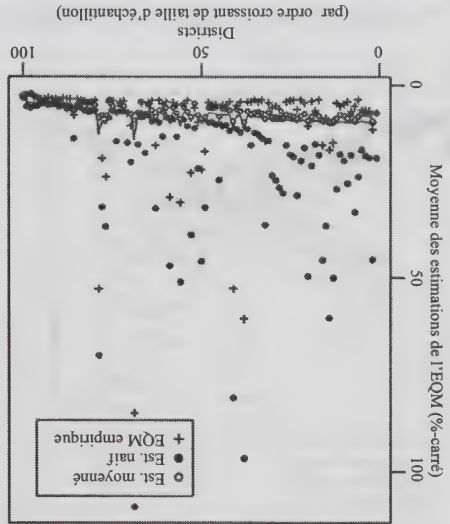


Figure 7 Moyenne et racine de l'EQM des estimateurs composites natif et moyen des EQM des pourcentages au niveau du district

En conclusion, cette simulation montre que, quand l'un des estimateurs de l'EQM, ici l'estimateur natif, est très inefficace, il contribue néanmoins, ne fût-ce que très modestement, à l'efficacité de l'estimateur composite de l'EQM. L'estimateur composite tire parti du meilleur des estimateurs moyen et natif, même dans des conditions non favorables. Une difficulté qui persiste est d'arriver à combiner les estimateurs natif et moyen de façon à satisfaire un critère particulier représentant un compromis entre le niveau de précision correspondant aux districts pour lesquels l'estimation est de haute précision et un plus haut niveau de précision pour les districts où la précision des estimations est faible. Par exemple, nous pourrions nous préoccuper moins de l'estimation de l'EQM pour les districts dont la représentation dans l'échantillon est importante et davantage de celle de l'EQM pour les districts représentés parcimonieusement. En outre, certains districts (par exemple, ceux situés dans une région particulière) pourraient présenter un intérêt spécial, non relié à leur représentation. Naturellement, dans ces circonstances, la première étape est la définition d'un critère ou d'une classe de critères qui relient les priorités inférentielles, et cette définition sera obligatoirement particulière à chaque enquête et à chaque client. Voir Longford (2006) pour certaines propositions.

4.1 Perfectionnements et extensions

Plusieurs éléments de réalisme peuvent être intégrés dans la dérivation de l'estimateur composite de l'EQM. Premièrement, l'incertitude au sujet de μ_d peut être reflétée en reconnaissant que μ_d et $\hat{\mu}_d$ sont corrélés. Donc, $\text{var}(\hat{\mu}_d - \mu) = \sigma_w^2(1/n_d - 1/n)$ et l'approximation (5) devient une égalité quand les deux occurrences de σ_w^2/n_d sont remplacées par $\sigma_w^2(1/n_d - 1/n)$. Cela n'entraîne qu'un léger changement quand $n_d \ll n$, ce qui est le cas pour la plupart des districts. Si le pays possède un district dominant, dont la taille d'échantillon représente une grande fraction de la taille d'échantillon globale, cet ajustement pourrait être pertinent, mais il a un effet négligeable sur l'estimation de l'EQM, car même l'estimation directe de la moyenne pour le district est quasi efficace. Un perfectionnement similaire peut être appliqué à l'estimateur empirique bayésien de μ_d . Il revient remplacer n_d par $1/(n_d^{-1} - n^{-1}) = n_d n / (n - n_d)$ dans le coefficient $b_d = 1/(1 + n_d \omega)$. Le changement ne devient non négligeable que dans le cas d'un district dominant, mais pour un tel district, le rééchantillonnage ne produit qu'une très petite amélioration par rapport à l'estimation directe avec ou sans cet ajustement. L'adaptation à des plans d'échantillonnage qui diffèrent de l'échantillonnage aléatoire stratifié et qui associent des sujets à des poids d'échantillonnage ne génère pas plus de

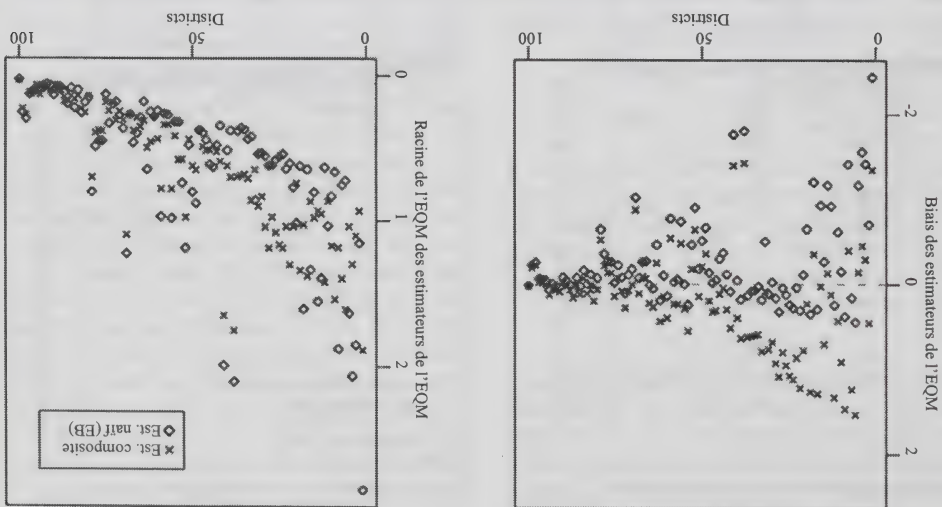


Figure 6 Biais et racine de l'EQM des estimateurs composite et naïf empirique bayésien de l'EQM de μ_d

district est de 6,85; l'asymétrie importante de ces pourcentages (coefficient d'asymétrie égal à 1,01 et aplatissement égal à 3,78) représente un test sévère de la méthode.

Dans la simulation, les pourcentages au niveau du district sont estimés par la version univariée de la méthode de régression (chapitre 8). Les résultats sont résumés à la figure 7. L'EQM est surestimée par les trois estimateurs pour la plupart des districts, sauf une minorité pour lesquels l'EQM empirique est plusieurs fois plus élevée que pour les autres. L'estimateur naïf présente un biais important pour la plupart des districts. L'estimateur moyen est contrôlé moins strictement que pour les résultats qui suivent une loi normale, car le coefficient de régression dépend aussi de la proportion estimée, qui est tronquée par le bas à 2 % pour éviter une variance estimée nulle $\hat{p}_d(1 - \hat{p}_d)/n_d$. Le graphique des estimations composées de l'EQM présente des pics pour les districts appropriés, mais ces pics sont beaucoup trop courts pour réduire considérablement le biais. Dans le cas de l'estimateur moyen, les EQM sont pour la plupart satisfaisantes, mais elles sont très grandes pour plusieurs districts. Pour ces districts, l'estimateur naïf de l'EQM est encore moins efficace que l'estimateur composite de l'EQM, compensés par les gains d'efficacité pour les faibles, mais le nombre de districts, mais les écarts sont assez faibles, pour lesquels l'estimateur moyen est moins efficace. L'estimateur naïf EB de l'EQM ressemble à de nombreux regards à l'estimateur naïf de l'EQM, il n'est pas représenté dans le diagramme.

Le manque d'efficacité de \hat{V}_d est dû, en partie, à son biais; ce dernier est supérieur à celui de \hat{V}_d^* pour tous les districts, sauf deux, mais la différence n'est non négligeable que si les deux estimateurs sont biaisés positivement. Donc, le gain d'efficacité est faible si l'on effectue l'analyse de façon à ce que les hypothèses concernant les lois soient satisfaites et faibles. Les gains sont modestes comparativement à l'accroissement de la difficulté à estimer l'efficacité, telle qu'elle est exprimée par $EQM(\hat{V}_d^*; V_d)$. Bien que la variance d'échantillonnage de $\hat{\sigma}_d^2$ soit négligeable dans les enquêtes à grande échelle, la contribution de $EQM(\mu_d; \mu_d)$ à $EQM(\hat{V}_d^*; V_d)$ ne peut pas être ignorée.

La figure 6 compare l'estimateur composite de l'EQM à l'estimateur naïf de \hat{V}_d basé sur l'estimateur empirique bayésien de μ_d . Il est dérivé par substitution de $\hat{\mu}_d$ à μ_d dans (2). Par souci de concision, nous l'appelons estimateur naïf EB. Comme prévu à la section 3, il a tendance à sous-estimer sa cible. Il est plus efficace que l'estimateur composite de l'EQM pour la moitié des districts (52 sur 100), mais ses propriétés sont moins uniformes. En principe, l'estimateur naïf EB pourrait être amélioré par combinaison à l'estimateur moyen; cependant, cette combinaison ne produit qu'une amélioration faible, même dans les conditions favorables (μ , σ_d^2 et σ_B^2 connues), et elle est nuisible pour plusieurs districts dans les conditions moins favorables. Nous omettons les détails.

À titre de simulation finale, nous considérons une variable de résultat binaire qui indique si $Y < 5$, de façon que les pourcentages au niveau du district soient dans la fourchette de 1,5 à 18,8 et que la dépendance du pourcentage à l'égard de la variance intra-district soit importante. La moyenne des pourcentages au niveau du

Figure 5 Biases et EQM des estimateurs de v_d . Les segments verticaux relient les valeurs associées à v_d^* et à v_d . Les valeurs associées à v_d^* sont représentées par des points noirs

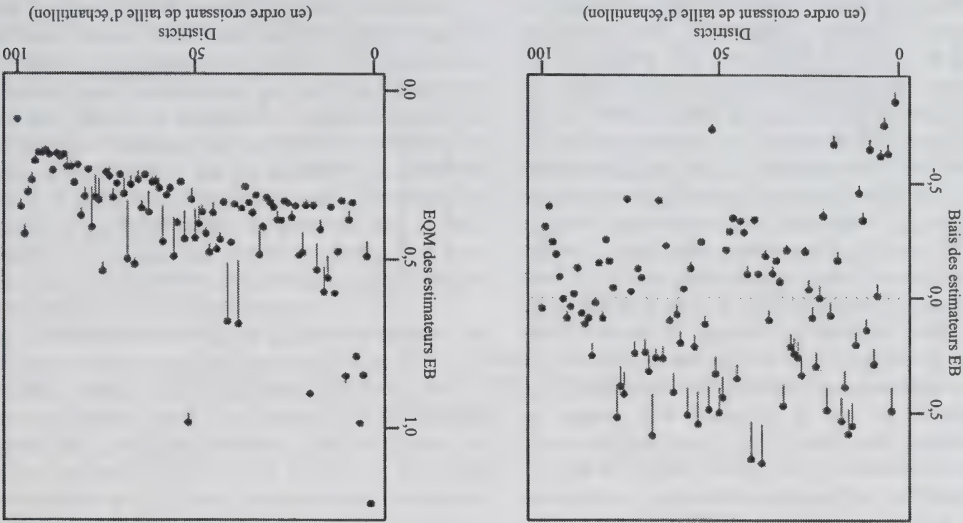
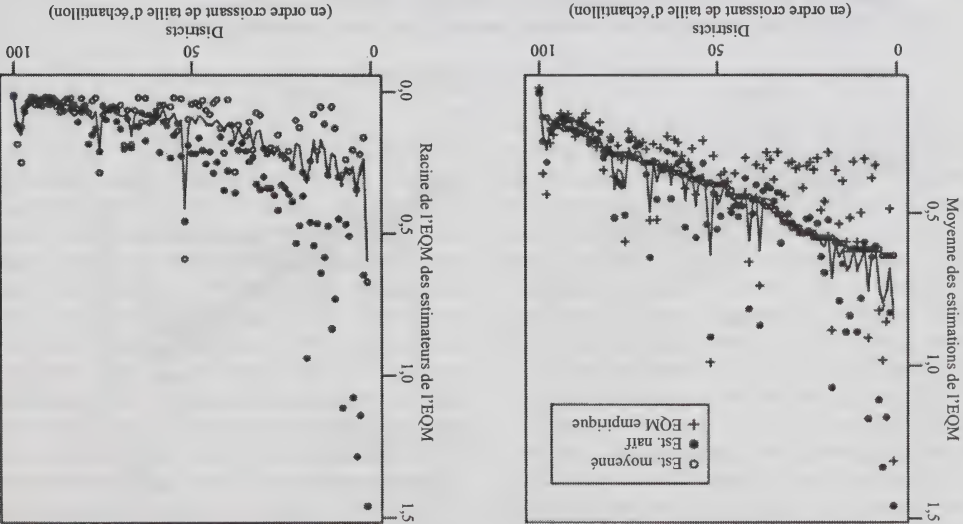


Figure 4 Moyenne et racine de l'EQM des estimateurs de v_d / 100 empiriques bayésiens; estimation des moyennes de v_d / 100



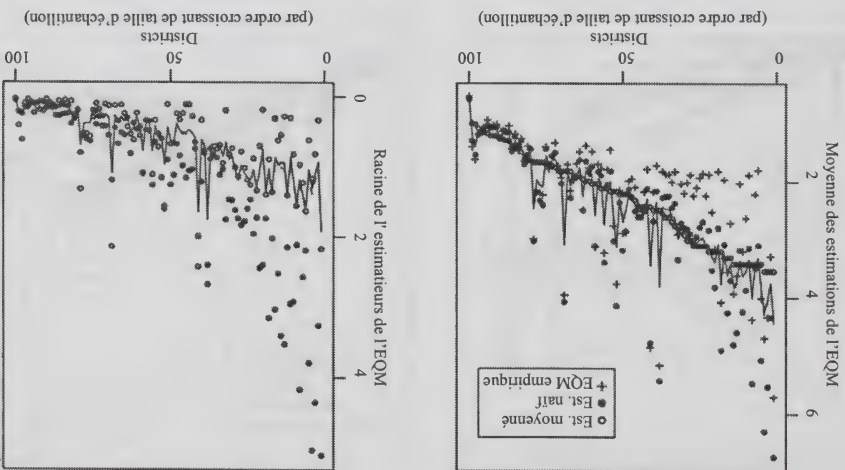


Figure 3 Moyenne et racine de l'EQM des estimateurs de l'EQM des paramètres globaux μ , σ_v^2 et σ_B^2 sont estimés (par ordre croissant de taille d'échantillon)

$$\hat{v}_d = \frac{100}{\hat{\mu}_d^2 - \widehat{\text{EQM}}(\hat{\mu}_d) + \hat{\sigma}_d^2} \quad (8)$$

Puis, nous comparons les estimateurs de l'EQM pour les moyennes de $Y^2/100$ au niveau du district que nous dénotons v_d . Les hypothèses de normalité intra et inter-districts ne sont plus appropriées. Nous appliquons les méthodes qui s'appuient sur les hypothèses de normalité pour évaluer la robustesse des estimateurs composites, mais aussi pour mettre en contraste les différences dues au « mauvais » choix de la transformation quadratique, parce que les espérances intra-district sont connues, égales à $(\mu_d^2 + \sigma_d^2)/100$, et pourraient être estimées par

Nous dénotons par \hat{v}_d les estimateurs empiriques bayésiens appliqués à $Y^2/100$. Les résultats des simulations fondées sur les valeurs de $Y^2/100$ sont présentés à la figure 4, en utilisant la même disposition et les mêmes symboles qu'à la figure 3. Nous arrivons aux mêmes conclusions qu'auparavant au sujet des biais et des racines de l'EQM, excepté que l'estimateur naïf est encore plus inefficace et que les propriétés de l'estimateur moyen sont encore plus erratiques, en ce sens qu'il est à la fois très efficace et très inefficace pour un plus grand nombre de districts que dans les conditions plus favorables de la figure 3. L'estimateur naïf est prudent, mais pour certains districts dont la taille d'échantillon n_d est faible, il l'est nettement trop, et son EQM est très grande. Nous contrastons ces conclusions à l'aide d'une comparaison de l'estimation des moyennes de $Y^2/100$ au niveau du district par \hat{v}_d , en transformant les estimations $\hat{\mu}_d$ conformément à (8). L'estimateur \hat{v}_d est plus efficace que \hat{v}_d pour la plupart des districts (90, en fait), et quand il est moins efficace, la différence relative entre leurs EQM est inférieure à 4 %. Pour quelques districts, la différence d'efficacité est perceptible, dépassant 20 % pour dix districts. Toutefois, les écarts entre les EQM sont faibles comparativement aux biais dans l'estimation de ces EQM, comme le montre la figure 5. Les biais et les EQM de \hat{v}_d sont représentés par des points noirs reliés entre eux pour

efficace, parce que l'estimateur naïf est très inefficace. Pour quelques-uns de ces districts, la combinaison des estimateurs est contreproductive, à cause du moyennage, mais il est impossible d'identifier ces districts d'après une seule réalisation de l'enquête.

Nous étudions maintenant des conditions moins favorables, sous lesquelles les hypothèses de normalité de μ_d sur l'ensemble des districts et des observations élémentaires y_{id} dans les districts sont encore satisfaites, mais les paramètres globaux, μ , σ_d^2 et σ_b^2 sont inconnus et estimés. Nous utilisons les mêmes moyennes μ_d et tailles d'échantillon n_d qu'à la figure 1. Les résultats des simulations sont résumés à la figure 3. Dans le panneau de gauche, nous représentons les moyennes empiriques des estimateurs de l'EQM en utilisant les mêmes symboles qu'à la figure 2, ainsi que les EQM empiriques (représentées par des croix

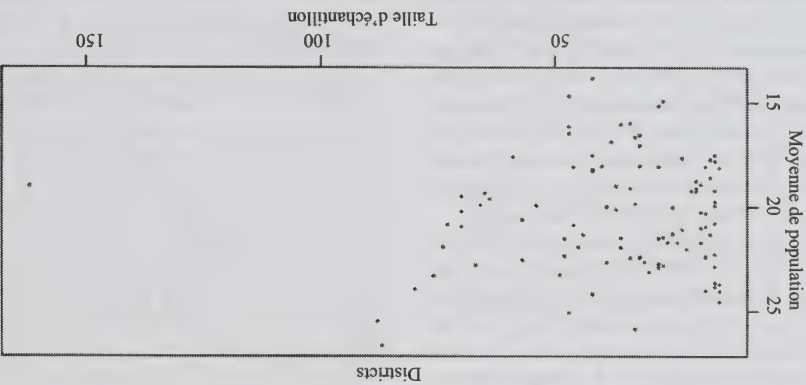


Figure 1 Valeurs d'échantillon au niveau du district et moyennes de population de X_i

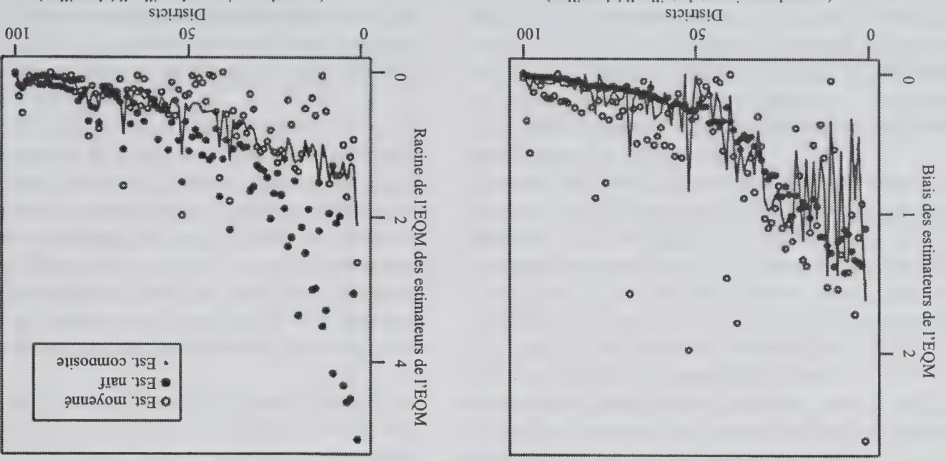


Figure 2 Biais et racine de l'EQM des estimateurs de l'EQM des estimateurs pour petits domaines empiriques bayésiens. Fondés sur des simulations dans des conditions artificielles. Les valeurs du biais et de la racine de l'EQM de l'estimateur composite sont reliées par des traits pleins (par ordre croissant de taille d'échantillon)

b_j . Dans les dérivations, nous avons utilisé l'identité $b_j = 1/(1 + n_j\omega)$, de sorte que cette expression ne pourrait pas être utilisée si les valeurs de b_j étaient fixées a priori.

4. Simulations

Les propriétés de l'estimateur composite de l'EQM ne pouvant être dérivées analytiquement, nous recourons à la simulation. Nous considérons les conditions artificielles d'une enquête nationale menée selon un plan d'échantillonnage stratifié, dont les strates coïncident avec les 100 districts du pays pour lesquels on souhaite estimer la moyenne d'une variable Y . Un échantillonnage aléatoire simple est appliqué à chaque strate, en supposant que la taille de leur population respective est pratiquement infinie. Nous avons généré les valeurs des moyennes à partir de la loi normale $N(\mu = 20, \sigma^2 = 8)$, et les tailles d'échantillon n_j , à partir de lois bêta conditionnelles normées, sachant les moyennes μ_j de façon à injecter un minimum de dépendance des moyennes à l'égard des tailles d'échantillon. Avec cet ajustement, l'hypothèse sur laquelle s'appuie l'estimateur moyen de l'EQM est faussée, mais cela risque de ne pas être décelé par une méthode diagnostique ou un test d'hypothèse, même dans le cas où μ_j est connue. La taille d'échantillon de l'un des districts a été modifiée de façon qu'elle soit beaucoup plus grande que les autres, afin de représenter la capitale du pays fictice. Les lois intra-strate de Y sont $N(\mu_j, \sigma_j^2 = 100)$. Les moyennes dans les répétitions. En guise d'orientation, elles sont représentées graphiquement à la figure 1. Des numéros d'ordre allant de 1 à 100 sont affectés aux districts par ordre croissant de taille d'échantillon. La taille d'échantillon la plus faible est $n_1 = 15$ et la taille d'échantillon globale est 3 698.

Dans les simulations, qui comprennent 1 000 répétitions, nous générerons les estimations directes $\hat{\mu}_j$ par des tirages aléatoires indépendants à partir de $N(\mu_j, \sigma_j^2/n_j)$ et les sommes corrigées des cartes intra-district, par des tirages indépendants à partir des lois du χ^2 normées appropriées avec $n_j - 1$ degrés de liberté. Puis, nous évaluons l'estimateur de rétrécissement $\hat{\mu}_j$ pour chaque district j , et ensuite, l'estimateur moyen $\hat{\mu}$, naïf et les deux estimateurs composites de l'EQM en utilisant les coefficients c_j^* et c_j^\dagger ou les estimations naïves.

Dans le premier ensemble de répétitions, nous supposons que μ_1, σ_1^2 et σ_2^2 sont connues, de sorte que la simulation reproduit les résultats dérivés théoriquement et nous permet d'évaluer la qualité des estimateurs composites de l'EQM sans l'interférence de l'incertitude au sujet du coefficient de rétrécissement $b_j = 1/(1 + n_j\omega)$. Les résultats sont résumés

graphiquement à la figure 2. Les biais empiriques (valeurs absolues) des quatre estimateurs de l'EQM sont tracés dans le panneau de gauche. Des cercles et des points noirs sont utilisés pour les estimateurs moyen et naïf, respectivement, et les biais des estimateurs composites sont reliés par des traits pleins. Les valeurs absolues des biais empiriques sont tracées afin de mettre en relief leur forte association à la taille d'échantillon dans le cas de l'estimateur naïf. Pour 60 districts (60 %), l'estimateur composite de l'EQM présente un biais positif. Pour l'estimateur naïf, ce chiffre ou pourcentage est plus élevé (78), et pour l'estimateur moyen, il est plus faible (52). Partout, le contributeur principal au biais de l'estimateur moyen de l'EQM est l'écart de la distance quadratique $(\mu_j - \mu)^2$ par rapport à la variance au niveau du district σ_j^2 . Les deux estimateurs composites, fondés sur $(\mu_j - \mu)^2$ et sur sa version corrigée du biais, diffèrent si peu qu'il est impossible de faire la distinction entre leurs biais dans le tracé. Le diagramme montre que l'estimateur moyen de l'EQM comporte un biais important pour quelques districts, y compris plusieurs dont la taille d'échantillon est grande. Les biais des estimateurs naïf et composite ne présentent pas de tels extrêmes.

Dans le panneau de droite, les racines de l'EQM des estimateurs de l'EQM sont représentées en utilisant les mêmes symboles et disposition. Le diagramme montre que l'estimateur naïf est inefficace, surtout pour les districts ayant les tailles d'échantillon les plus faibles, tandis que l'estimateur moyen est très efficace pour certains, mais inefficace pour d'autres, sans aucune relation apparente avec leur taille d'échantillon. En fait, mise à part la taille d'échantillon, la grande efficacité est associée à la proximité de $(\mu_j - \mu)^2$ par rapport à σ_j^2 et la faible efficacité, aux valeurs les plus petites et les plus grandes de $(\mu_j - \mu)^2$. Par exemple, la racine de l'EQM empirique de l'estimateur moyen de l'EQM pour le district 1, avec $n_1 = 15$, est 2,63, tandis que son équivalent pour le district 11 ($n_{11} = 16$) est 0,049. Les moyennes de population sont $\mu_1 = 24,55$, $\mu_{11} = 22,87$, et $\sigma_1^2 = 1,72$, et $\sigma_{11}^2 = 0,40$. Les racines de l'EQM pour l'estimateur naïf sont 5,08 et 3,51, et celles pour l'estimateur composite, 2,10 et 1,00 pour les districts 1 et 11, respectivement. L'estimateur composite de l'EQM donne des résultats nettement plus uniformes, atténuant les défauts des estimateurs moyen et naïf.

Les trois estimateurs sont prudents (ont un biais positif) pour les districts pour lesquels l'EQM de μ_j est relativement faible. L'estimateur moyen possède un biais négatif lorsque les EQM sont relativement grandes. L'estimateur composite est également enlaid d'un biais négatif pour certains districts, mais celui-ci a tendance à être plus faible qu'il ne l'est pour les districts ayant les tailles d'échantillon les plus faibles, l'estimateur composite n'est pas très

sur l'ensemble des districts d . Nous remplaçons $(\mu_d - \mu)^2$ par σ_B^2 , et $(\mu_d - \mu)^4$ par $3\sigma_B^4$ ou, en général, par $k\sigma_B^k$, où k est le degré de la distribution (au niveau du district) de μ_d . Bien qu'il puisse sembler, à première vue, que nous n'ayons rien gagné, parce que nous devons encore éliminer la dépendance de l'EQM à l'égard de $(\mu_d - \mu)^2$ en utilisant σ_B^2 à la place, nous procédons maintenant à cette étape à un stade ultérieur. Dans les simulations présentées à la section 4, nous montrons que cela réduit l'effet indésirable du calcul d'une moyenne, ou moyennage.

Donc, nous recherchons le coefficient c_d qui minimise l'EQM espérée de l'estimateur composite de l'EQM.

$$\begin{aligned} \widehat{\text{EQM}}(\hat{\mu}_d; \mu_d) &= (1 - c_d) \widehat{\text{EQM}}(\hat{\mu}_d; \mu_d) + c_d \widehat{\text{EQM}}(\hat{\mu}_d; \mu_d) \\ &= (1 - c_d) \left\{ (1 - b_d)^2 \frac{n_d}{\sigma_B^2} + b_d^2 (\mu_d - \mu)^2 \right\} + c_d b_d^2 \sigma_B^2. \quad (7) \end{aligned}$$

Pour évaluer l'EQM de l'estimateur de l'EQM, sous la forme d'une fonction de c_d , nous utilisons les expressions

$$\begin{aligned} \overline{\text{MSE}}\{b_d^2 \sigma_B^2; \text{MSE}(\hat{\mu}_d; \mu_d)\} &= 2b_d^4 \sigma_B^4, \\ \overline{\text{MSE}}\{(\hat{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} &= \frac{n_d^2}{\sigma_B^4} (3 + 4n_d \omega), \\ \overline{\text{MSE}}\{(\hat{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} &= \frac{n_d^2}{\sigma_B^4} \{3(1 - b_d)^4 + 3b_d^2(2 - b_d)^2 n_d^2 \omega^2 \\ &\quad + 2(1 - b_d)^2(2 - b_d + 3b_d^2)n_d \omega\}, \end{aligned}$$

obtenues en calculant la moyenne des équations respectives 4), 5) et 6); $(\mu_d - \mu)^2$ est remplacé par σ_B^2 et $(\mu_d - \mu)^4$ par $3\sigma_B^4$. En supposant que les cibles au niveau du district μ_d suivent une loi normale, l'EQM de l'estimateur composite (7) est

$$\begin{aligned} E\{(1 - c_d)(1 - b_d)^2 \frac{n_d}{\sigma_B^2} + (1 - c_d)b_d^2(\mu_d - \mu)^2\} &= b_d^2 E\{(1 - c_d)(\mu_d - \mu)^2 + c_d \sigma_B^2\} \\ &\quad + c_d \sigma_B^2 (1 + n_d \omega - \omega^2 n_d^2 \omega - (\mu_d - \mu)^2)^2 \\ &= b_d^2 E\{(1 - c_d)\sigma_B^2 n_d^2 \omega + (1 - c_d)(\mu_d - \mu)^2 \\ &\quad + c_d b_d^2 \sigma_B^2 - b_d^2 \sigma_B^2 n_d^2 \omega^2 - b_d^2 (\mu_d - \mu)^2\}^2 \\ &\quad + c_d \sigma_B^2 \left\{ 2\sigma_B^4 \frac{n_d^2}{\omega} + 4\sigma_B^2 (\mu_d - \mu)^2 \right\} \\ &\quad + b_d^2 \left\{ (1 - c_d) \frac{n_d}{\sigma_B^2} + c_d \sigma_B^2 - (\mu_d - \mu)^2 \right\}^2, \end{aligned}$$

en utilisant les identités $(1 - b_d)^2 = b_d^2 n_d^2 \omega^2$ et $\sigma_B^2 = \sigma_W^2 \omega$ pour extraire le facteur b_d^4 . En calculant l'espérance sur l'ensemble des districts en gardant intactes les tailles d'échantillons, nous obtenons

$$\widehat{\text{EQM}}\{\widehat{\text{EQM}}(\hat{\mu}_d; \mu_d)\} = \frac{b_d^4}{n_d^2} \{ (1 - c_d)^2 (3 + 4n_d \omega) \sigma_W^4 + 2c_d^2 n_d^2 \sigma_B^4 \}.$$

Le minimum de cette fonction quadratique de c_d est atteint pour

$$c_d^* = \frac{3 + 4n_d \omega}{3 + 4n_d \omega + 2n_d^2 \omega^2}.$$

Ce choix d'un coefficient c_d concorde avec nos attentes. Pour $n_d = 0$, $c_d^* = 1$ et nous nous appuyons uniquement sur l'estimateur de l'EQM moyen, égal à σ_B^2 . En outre, c_d^* est une fonction décroissante de n_d , qui converge vers zéro à mesure que n_d diverge vers $+\infty$; pour les grandes valeurs de n_d , nous utilisons l'estimateur naïf de l'EQM. Il s'agit également d'une fonction décroissante de ω ; pour $\omega = 0$, c'est-à-dire $\sigma_B^2 = 0$, $c_d^* = 1$ pour chaque district d , ce qui confirme que $\mu_d \equiv \mu$ et que μ_d serait estimée avec précision si μ était connue. À mesure que ω augmente, $\sigma_B^2/(1 + n_d \omega)$ devient de moins en moins utile, parce que les écarts quadratiques $(\mu_d - \mu)^2$ sont fortement étalés (autour de σ_B^2).

Si nous corrigeons $(\mu_d - \mu)^2$ de son biais en estimant l'EQM espérée de l'estimateur de rétrécissement $(\mu_d - \mu)^2$, l'EQM est minimisée pour

$$c_d^* = \frac{1 + 2n_d \omega}{1 + n_d \omega^2}.$$

Il est facile de vérifier que

$$c_d^* - c_d^* = \frac{(1 + n_d \omega)^2}{n_d^2 \omega^2} \frac{3 + 4n_d \omega + 2n_d^2 \omega^2}{1},$$

de sorte qu'il soit assigné à l'estimateur corrigé du biais dérivé de (2) un poids plus grand (égal à $1 - c_d^*$) qu'à l'estimateur naïf. Toutefois, la différence est faible pour toutes les valeurs de $n_d \omega$.

L'estimateur composite de l'EQM basé sur $(\mu_d - \mu)^2$ s'obtient de façon similaire, mais l'expression résultante est nettement plus compliquée. Le coefficient de rétrécissement optimal est

$$\begin{aligned} c_d^* &= 3(1 - b_d)^4 + 2(1 - b_d)^2 f(b_d) n_d \omega - b_d^2 (2 - b_d) f(b_d) n_d^2 \omega^2 \\ &\quad \times \{3(1 - b_d)^4 + 2(1 - b_d)^2 f(b_d) n_d \omega - \\ &\quad \{2 - 4b_d(2 - b_d) + 3b_d^2 f(b_d)\} n_d^2 \omega^2\}. \end{aligned}$$

ou $f(b_d) = 2 - 6b_d + 3b_d^2$. La dépendance à l'égard de b_d est particulièrement préoccupante, car en pratique, b_d est estimé et les propriétés de l'estimateur de l'EQM fondé sur le coefficient c_d^* estimé sont obligatoirement affectées par

chacune d'elle peut varier. EQM peut être interprétée comme l'espérance du modèle, quoique l'espérance ou la moyenne des écarts quadratiques $(\mu_d - \mu)^2$ puisse être considérée et estimée pour un ensemble donné de districts sans aucune référence à un modèle. Dans (3), la variance conditionnelle est appropriée pour les districts pour lesquels μ_d est dans la distance « typique », σ_B , par rapport à la moyenne nationale μ . Si $|\mu_d - \mu| \neq \sigma_B$, un estimateur sans biais de la variance conditionnelle $\sigma_B^2/(1+n_d\omega)$ est biaisé pour $EQM(\bar{\mu}_d; \mu_d)$. Comme le biais est relié à la grandeur de la population $\mu_d - \mu$, il n'est pas réduit par l'accroissement de la taille d'échantillon n_d .

3. Estimations composites de l'EQM

Pour estimer $EQM(\bar{\mu}_d; \mu_d)$, nous réunissons l'idée du rétrocissement et combinons deux estimateurs possibles, c'est-à-dire $\sigma_B^2/(1+n_d\omega)$ et un estimateur naïf de l'EQM donné par (2). Cet estimateur composite peut être justifié comme il suit. Si $n_d = 0$, et par conséquent $\bar{\mu}_d = \mu$, nous ne possédons aucune information directe au sujet de μ_d , de sorte que nous ne pouvons pas améliorer $\sigma_B^2/(1+n_d\omega)$ en tant qu'estimateur de $EQM(\bar{\mu}_d; \mu_d)$. Quand n_d est grand, μ_d est estimée avec une précision suffisante pour utiliser $(\bar{\mu}_d - \mu)^2$, éventuellement avec une correction du biais, comme estimateur de $(\mu_d - \mu)^2$. Pour les tailles d'échantillon intermédiaires, nous recherchons une combinaison (compromis) de ces deux alternatives qui sont appropriées dans les conditions extrêmes, c'est-à-dire quand $n_d = 0$ et quand $n_d \rightarrow +\infty$. Par conséquent, nous devons des expressions pour leurs EQM, puis pour l'EQM de leur

$$EQM\left\{\frac{\sigma_B^2}{1+n_d\omega}; EQM(\bar{\mu}_d; \mu_d)\right\} = \left\{\frac{\sigma_B^2}{1+n_d\omega} - \frac{\sigma_B^2}{1+n_d\omega} \frac{1}{1+n_d\omega} - \frac{\sigma_B^2}{1+n_d\omega} \frac{1}{1+n_d\omega} \frac{1}{1+n_d\omega} \right\} \quad (4)$$

L'écart quadratique $(\mu_d - \mu)^2$, qui intervient dans (2), est estimé naïvement par $(\bar{\mu}_d - \mu)^2$ avec un biais égal à σ_B^2/n_d et, en supposant que $\bar{\mu}_d$ suit une loi normale,

$$EQM\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} = \text{var}\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} + [E\{(\bar{\mu}_d - \mu)^2 - (\mu_d - \mu)^2\}]^2 \\ = 2\sigma_d^4 \frac{\sigma_B^2}{\sigma_d^2} + 4(\mu_d - \mu)^2 \frac{\sigma_B^2}{\sigma_d^2} + \frac{n_d}{2} \frac{\sigma_B^2}{\sigma_d^2} + \left\{ \frac{\sigma_B^2}{3\sigma_d^2} + \frac{n_d}{4(\mu_d - \mu)^2} \right\} \quad (5)$$

dérivée d'après les propriétés de la loi de χ^2 non centré et d'une approximation en permettant que $n \rightarrow +\infty$. Une autre option consiste à utiliser $\bar{\mu}_d$ au lieu de μ_d ; des opérations élémentaires donnent les approximations

$$E\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} = (1 - b_d)^2 \left\{ \frac{\sigma_B^2}{\sigma_d^2} + (\mu_d - \mu)^2 \right\} \\ \text{var}\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} = \frac{n_d^2}{(1 - b_d)^4} \sigma_B^2 (2\sigma_d^2 + 4n_d(\mu_d - \mu)^2), \\ \text{où } b_d = 1/(1 + n_d\omega), \text{ et donc} \\ EQM\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} = \text{var}\{(\bar{\mu}_d - \mu)^2; (\mu_d - \mu)^2\} + [E\{(\bar{\mu}_d - \mu)^2 - (\mu_d - \mu)^2\}]^2 \\ = (1 - b_d)^4 \frac{n_d^2}{3\sigma_d^4} + 2(1 - b_d)^2 (2 - b_d^2) (\mu_d - \mu)^2 + 2(1 - b_d)^2 \frac{\sigma_B^2}{\sigma_d^2} (\mu_d - \mu)^2 \quad (6)$$

Cette approximation est valide uniquement pour $b_d = 1/(1 + n_d\omega)$, de sorte qu'une approximation supplémentaire intervient quand nous substituons un choix éventuellement sous-optimal ou une estimation de b_d fondée sur une estimation de ω . En général, le coefficient b_d qui minimise l'EQM dans (6) diffère de $1/(1 + n_d\omega)$ parce qu'avec $b_d = 1/(1 + n_d\omega)$, le rétrocissement est optimal uniquement pour les cibles qui sont des transformations linéaires de μ_d (Shen et Louis 1998). Nous ne poursuivons pas cette route car, étant une fonction compliquée des paramètres, la solution est vraisemblablement sensible à l'erreur dans l'estimation de ces derniers. L'estimateur $(\bar{\mu}_d - \mu)^2$ pourrait être corrigé de son biais en estimant $(\mu_d - \mu)^2$, quoique l'on risque d'obtenir une estimation négative, surtout si n_d est faible.

Enfin, nous combinons les deux estimateurs (biaisés) de $EQM(\bar{\mu}_d; \mu_d)$, l'estimateur moyené $\sigma_B^2/(1 + n_d\omega)$ et l'estimateur naïf dérivé de l'identité (2), en utilisant $(\bar{\mu}_d - \mu)^2$ comme estimateur de $(\mu_d - \mu)^2$. Les EQM de ces deux estimateurs dépendent de $(\mu_d - \mu)^2$, de sorte que nous remplaçons les termes pertinents par leurs espérances

Par contre, les méthodes fondées sur un modèle sont associées d'un beaucoup plus grand nombre d'hypothèses qui, souvent, ne peuvent être vérifiées. Diverses méthodes de diagnostic en vue d'évaluer la qualité de la modélisation sont disponibles, mais elles sont toutes teintées d'une incertitude. Interpréter la non-découverte d'une contradiction comme une preuve de l'absence de toute contradiction est une incohérence logique commise fréquemment. Elle ne peut être évitée qu'en mentionnant les propriétés des estimateurs quand les hypothèses ne sont pas valides, mais les méthodes de ce genre sont difficiles à élaborer, à cause de la vaste gamme de violations du modèle dont il faudrait tenir compte. Pourtant, malgré ces inconvénients, il s'est avéré que les méthodes d'estimation pour petits domaines fondées sur un modèle ont du mérite et sont aujourd'hui considérées, à juste titre, comme indispensables (Ghosh et Rao 1994; Rao 2003; Longford 2005).

Le projet EURAREA (EURAREA Consortium 2004) a mené une étude par simulation à grande échelle comportant l'échantillonnage de populations artificiellement générées ressemblant aux populations humaines de plusieurs pays européens et l'application de plusieurs classes d'estimateurs. L'étude a confirmé la supériorité des estimateurs fondés sur un modèle, avec plusieurs réserves, mais a produit des résultats plutôt décevants en ce qui concerne les estimateurs de leurs erreurs-types. Nous imputons le problème à l'application d'un calcul de moyenne dans la dérivation des erreurs-types des estimateurs de rétrécissement.

Supposons qu'une population est divisée en D districts, chacun ayant une taille de population pouvant, à toutes fins utiles, être considérée infinie, et que des plans d'échantillonnage aléatoire simple indépendants sont appliqués dans les districts. Nous supposons que, dans chaque district d , la variable de résultat Y suit une loi normale de moyenne μ_d et de même variance σ_w^2 , $N(\mu_d, \sigma_w^2)$. Pour les moyennes de population intra-district μ_d , nous supposons le modèle de population $\mu_d \sim N(\mu, \sigma_B^2)$, mais nous voulons faire des inférences au sujet d'un ensemble fixe de moyennes (réalisées) $\{\mu_d\}$. À la section 5, nous discutons des conditions de régression plus générales définies par les modèles intra-district

$$Y|d \sim N(X_d \beta + \delta_d, \sigma_w^2),$$

dans lesquels X_d sont les matrices de régression intra-district, β est l'ensemble de paramètres de régression correspondants communs à tous les districts, et δ_d est l'écart de la régression intra-district par rapport à la régression typique définie par $\delta_d = 0$. Dans la super-population, δ_d représentent un échantillon aléatoire tiré d'une loi $N(0, \sigma_\delta^2)$, mais nous voulons faire des inférences au sujet de l'ensemble fixe (réalisé) $\{\delta_d\}$. Donc, nous utilisons des estimateurs fondés sur un modèle, mais nous

évaluons leurs propriétés en fonction de critères fondés sur le plan. Dénotons par μ la moyenne (nationale) des grands domaines μ_d et par σ_B^2 la variance entre les districts, $\sigma_B^2 = D^{-1} \sum_d (\mu_d - \mu)^2$. Notons que ces paramètres diffèrent de leurs équivalents en superpopulation μ et σ_B^2 . Nous supposons pour commencer que σ_B^2, σ_w^2 et μ sont connues. Soit μ_d et μ les moyennes d'échantillon de la variable d'intérêt dans le district d et dans le domaine complet (pays). Elles sont fondées sur des échantillons de tailles n_d et $n = n_1 + \dots + n_D$. Si l'on n'utilise aucune covariable, l'estimateur empirique bayésien (de rétrécissement) de μ_d est

$$\hat{\mu}_d = \left(1 - \frac{1 + n_d \omega}{1 + n \omega}\right) \mu_d + \frac{1 + n_d \omega}{1 + n \omega} \mu, \quad (1)$$

où $\omega = \sigma_B^2 / \sigma_w^2$ est le ratio des variances. La variance conditionnelle fondée sur le modèle de μ_d , sachant les données, μ , σ_B^2 et σ_w^2 (égale à $\sigma_B^2(1 + n_d \omega)$, est souvent considérée comme la variance d'échantillonnage de $\hat{\mu}_d$. Les origines de cette pratique remontent à l'application de l'algorithme EM. Une dérivation plus minutieuse tient compte du fait que, dans l'approche fondée sur le plan de sondage, μ_d est biaisé pour μ_d .

$$E(\hat{\mu}_d | \mu_d) - \mu_d = -\frac{1 + n_d \omega}{1 + n \omega},$$

et que sont erreur quadratique moyenne est

$$EOM(\hat{\mu}_d; \mu_d) = \left(1 - \frac{1 + n_d \omega}{1 + n \omega}\right)^2 \text{var}(\mu_d) + \frac{(\mu_d - \mu)^2}{1 + n \omega} + \frac{\sigma_w^2}{n_d \omega^2} \left(\frac{1}{(\mu_d - \mu)^2} + \frac{1 + n_d \omega}{1 + n \omega} \right), \quad (2)$$

en supposant, pour simplifier, que $\mu = \mu$. Afin de souligner que l'EOM dépend de la cible, nous incluons l'estimateur ainsi que la cible dans son argument. En particulier, $EOM(\hat{\mu}_d; \mu_d) \neq EOM(\hat{\mu}_d; \mu_d)$, à moins que $\mu_d = \mu$. Une caractéristique gênante de l'identité (2) est qu'elle contient μ_d , la cible de l'estimation. Si nous remplaçons $(\mu_d - \mu)^2$ par son espérance sur l'ensemble des districts, σ_B^2 , nous obtenons l'identité plus connue

$$EOM(\hat{\mu}_d; \mu_d) = \frac{\sigma_B^2}{1 + n_d \omega}, \quad (3)$$

la variance fondée sur un modèle conditionnel reliée à EM de μ_d . La barre au-dessus de EOM indique l'espérance (moyenne) de $(\mu_d - \mu)^2$, le numérateur du dernier terme de (2), sur l'ensemble des districts, en gardant les tailles d'échantillon n_d intactes. Tout au long de l'exposé, nous conditionnons sur les tailles d'échantillon intra-district $n_d, d = 1, \dots, D$, même si, dans le plan d'échantillonnage,

De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle

Nicholas Tibor Longford¹

Résumé

Nous déterminons un estimateur de l'erreur quadratique moyenne (EQM) de l'estimateur de Bayes empirique et composite de la moyenne locale dans les conditions standard de petits domaines. L'estimateur de l'EQM est un composite de l'estimateur établi, fondé sur l'espérance conditionnelle de l'écart aléatoire associé au domaine, et d'un estimateur naïf de l'EQM fondé sur le plan de sondage. Nous évaluons ses propriétés par simulation. Enfin, nous examinons des variantes de cet estimateur de l'EQM et décrivons certaines extensions.

Mots clés : Estimation composite; estimation empirique bayésienne; rétrécissement; estimation pour petits domaines.

1. Introduction

Au fil des ans, les méthodes fondées sur le plan de

sondage se sont avérées inefficaces pour l'estimation pour petits domaines, parce que, contrairement aux méthodes empiriques bayésiennes et connexes, elles ne permettent pas d'utiliser efficacement l'information auxiliaire. Cependant, les hypothèses associées aux modèles utilisés demeurent une faiblesse des méthodes fondées sur un modèle, parce que les inférences qui en découlent ont le défaut généralisé de dépendre de la validité du modèle. Dans l'application des modèles empiriques bayésiens à l'estimation pour petits domaines, les zones locales (districts) sont associées à des effets aléatoires. Sous l'approche fondée sur le plan de sondage, cette hypothèse n'est pas valide, car, lors d'une répétition hypothétique de l'enquête, les mêmes districts seraient réalisés (à l'exception de certains qui ne sont pas représentés dans l'échantillon tiré), et les grands cibles associées à ces districts seraient également les mêmes. Autrement dit, les districts devraient être associés à des effets fixes. Le manque de validité de cet aspect des modèles empiriques bayésiens n'a aucun effet indésirable sur l'estimation des grands pour petits domaines (moyennes, totaux, proportions, etc.). L'association des petits domaines à des effets aléatoires est un élément essentiel à l'emprunt d'information aux autres domaines ou à l'exploitation des similarités entre les domaines, ainsi qu'entre les variables, les points dans le temps, les enquêtes et d'autres sources de données, mais elle fausse l'évaluation de la précision des estimateurs. Certains estimateurs composites et les estimateurs de leur erreur quadratique moyenne ont le même défaut.

À la section suivante, nous diagnostiquons le problème en détail et à la section 3, nous proposons une solution, que nous illustrons et évaluons ensuite à la section 4 par simulation en utilisant une série d'exemples. Ceux-ci varient du plus simple et favorable (en accord avec la plupart des hypothèses formulées) au plus complexe et le moins

favorable, afin d'explorer la robustesse de la méthode. Nous discutons de son potentiel de manière plus complète à la

2. Fixe et aléatoire

Par variance d'échantillonnage d'un estimateur général $\hat{\theta}$ fondé sur un processus de génération de données (échantillonnage) donné χ , nous entendons la variation des valeurs de $\theta(\mathbf{X})$ dans les répétitions des processus qui génèrent les ensembles de données \mathbf{X} et qui leur appliquent $\hat{\theta}$. Dans l'approche fondée sur le plan de sondage, la répétition d'une enquête à l'échelle d'un pays et sa division en D districts produit les mêmes quantités de population au niveau du district $\theta_d, d = 1, \dots, D$; ces D quantités sont fixes. En revanche, dans l'approche fondée sur un modèle, chaque répétition en utilisant un modèle empirique bayésien débute par la génération d'un nouvel ensemble de D valeurs de θ_d , indépendamment des répétitions antérieures.

Nous considérons l'approche fondée sur le plan comme appropriée, parce qu'en principe, chaque quantité θ_d pourrait être établie avec précision et qu'une répétition hypothétique de l'enquête correspondrait au tirage d'un échantillon à partir de la même population, avec la même division du pays en ses districts et les mêmes valeurs des variables enregistrées pour chaque membre de la population. La plupart des méthodes fondées sur le plan de sondage établies sont valides quand l'enquête porte sur une base de sondage parfaite, qui ne contient aucun enregistrement en double et s'applique exclusivement à la population étudiée, et que le plan d'échantillonnage est exécuté parfaitement, sans aucun écart par rapport au protocole établi. Autrement dit, les estimateurs qu'elles produisent sont (approximativement) sans biais, les expressions pour les variances d'échantillonnage sont correctes, ou le sont quasiment, et ces variances sont estimées avec un biais faible ou nul.

Kim, J.-K., Navarro, A. et Fuller, W. A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Lee, H., et Kim, J.-K. (2002). Jackknife variance estimation for two-phase samples with high sampling fractions. *Proceedings of ASA Section on Survey Research Methods*, 2024-2028.

Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data. *Biometrika*, 77, 365-375.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., et Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K., et Sitter, R.R. (1997). Variance estimation under stratified two-phase sampling with applications to measurement bias. Dans *Survey Measurement and Process Quality: Wiley Series in Probability and Statistics*. (Eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz et D. Trewin). New York: 753-768.

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Royall, R.M., et Cumberland, W.G. (1981b). The finite population linear regression estimator: An empirical study. *Journal of the American Statistical Association*, 76, 924-930.

Schreuder, H.T., Li, H.G. et Scott, C.T. (1987). Jackknife and bootstrap estimation for sampling with partial replacement. *Forest Science*, 33, 676-689.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag: New York.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Woodruff, R.S. (1952). Confidence intervals for median and other position measures. *Journal of the American Statistical Association*, 47, 635-646.

Wu, C., et Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *Journal of Official Statistics*, 19, 119-131.

Funakoka, F., Saigo, H., Sitter, R.R. et Toida, T. (2006). Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*, 32, 169-175.

Dennari, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Cochran, W.G. (1977). *Sampling Techniques*. 3^{ème} Edition. New York : John Wiley & Sons, Inc.

Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Biemer, P.P., et Atkinson, D. (1993). Estimation de l'erreur systématique de mesure par la prédiction modeliste. *Techniques d'enquête*, 19, 137-146.

Berger, Y.G., et Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 68, 531-547.

Bibliographie

Berger, Y.G., et Rao, J.N.K. (2006). Adjusted jackknife for imputation under probability sampling without replacement. *Journal of the Royal Statistical Society, B*, 68, 531-547.

Biemer, P.P., et Atkinson, D. (1993). Estimation de l'erreur systématique de mesure par la prédiction modeliste. *Techniques d'enquête*, 19, 137-146.

Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chen, J., Sitter, R.R. et Wu, C. (2002). Using empirical likelihood method to obtain range restricted weights in regression estimator for surveys. *Biometrika*, 89, 230-237.

Chen, J., et Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12, 1223-1239.

Cochran, W.G. (1977). *Sampling Techniques*. 3^{ème} Edition. New York : John Wiley & Sons, Inc.

Dennari, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Funakoka, F., Saigo, H., Sitter, R.R. et Toida, T. (2006). Bootstrap de type Bernoulli pour l'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*, 32, 169-175.

$$1. \quad \sum_{i=1}^n \bar{x}_i = E_{(A^*, B^*)}(\bar{x}^*) + V_{(A^*, B^*)}(E_{(A^*, B^*)}(\bar{x}^*)), \text{ où } V_{(A^*, B^*)} \text{ et } E_{(A^*, B^*)} \text{ sont, respectivement, la variance par rapport à l'échantillonnage } A^* \text{ et la variance conditionnelle par rapport à l'échantillonnage } B^* \text{ sachant } A^*,$$

où $\sum_{(j), \underline{x}^j, \underline{x}^j} (j, \underline{x}^j, \underline{x}^j)$ est la matrice de variance-covariance de l'échantillonnage bootstrap proposé. L'estimation convergente de la variance sous la méthode proposée est prouvée en montrant que Δh^j et $\sum_{(j), \underline{x}^j, \underline{x}^j} (\Delta h^j)^2$ respectivement convergent pour Δh et $\sum_{(j), \underline{x}^j, \underline{x}^j} (\Delta h)^2$ pour Δh pour Δh découle de la convergence de $(j, \underline{x}^j, \underline{x}^j)$ pour $(j, \underline{x}^j, \underline{x}^j)$ et de la continuité de h . La convergence de $\sum_{(j), \underline{x}^j, \underline{x}^j} (\Delta h^j)^2$ peut être montrée comme il suit. Pour commencer, puisque nous utilisons une méthode bootstrap appropiée pour l'échantillonnage aléatoire simple sans remise dans le sous-échantillonnage \mathcal{A}^j , nous avons $\sum_{(j), \underline{x}^j, \underline{x}^j} (j, \underline{x}^j, \underline{x}^j) = (1 - f) f^2 / v^j$, où $\sum_{(j), \underline{x}^j, \underline{x}^j} (j, \underline{x}^j, \underline{x}^j)$ avec $n = n^j$, $(j, \underline{x}^j, \underline{x}^j)$. Deuxièmement parce que

$$y_{\Delta}^{(g, \underline{x}^{(g)}, \underline{x}^{(g)}, \underline{x}^{(g)})} \sum y_{\Delta} = (\theta)^* A$$

$$\dot{\theta} = \dot{\theta} + \nabla h^* \left((\Delta \hat{x}^*)^{\gamma_1}, \Delta \hat{x}^{\gamma_2}, \Delta \hat{x}^{\gamma_3} \right) + o \left(\left\| u \right\|^{1/2} \right),$$

où $E_{\cdot, \cdot}^{\cdot, \cdot}$ et $E_{\cdot, \cdot}^{*, \cdot}$ sont, respectivement, l'espérance par rapport à l'échantillonnage \mathcal{A} et l'espérance conditionnelle sachant \mathcal{A}^* sous la méthode bootstrap proposée. Alors, $\hat{\theta}^* = h(\hat{f}^{\cdot, \cdot}, \hat{x}^{\cdot, \cdot}, \hat{x}^{\cdot, \cdot, \cdot})$ est approximé par

$$({}^v f - 1) / (({}^v \underline{x} - {}^v \underline{x}) {}^v f + {}^g \underline{x}) {}^y \underline{f} = {}^g \underline{x} =$$

$$E_{\cdot}(\underline{x}_{B_{\cdot}}) = E_{\cdot A_{\cdot}}(E_{\cdot B_{\cdot}|A_{\cdot}}(\underline{x}_{B_{\cdot}}))$$

Maintenant, considérons un développement en série de Taylor de $\hat{\theta} = h(\bar{x}^y, \bar{x}^x, \bar{x}^r)$ avec $\bar{\theta} = \bar{x}^y + f^y(\bar{x}^y, \bar{x}^x)$. Soit E_r et V_r l'espérance et la variance sous la procédure bootstrap proposée, respectivement. Pour commencer, observons que $E_r(\bar{y}^y) = \bar{y}^y$, $E_r(\bar{x}^x) = \bar{x}^x$ et

Saigo : Bootstrap

$$\begin{aligned} & \left(\nu_{\underline{A}} - \iota_{\mathcal{A}} \right) \left({}^{07}\zeta_{\underline{Z}} - \iota_{\mathcal{Z}} x \right) \nu_{\mathfrak{Z}!} \sum_{\perp} (1 - \nu u) \nu_{\mathcal{C}} = {}^{27}\lambda \\ & \left(\nu_{\underline{X}} - \iota_{\mathcal{X}} \right) \left({}^{07}\zeta_{\underline{Z}} - \iota_{\mathcal{Z}} x \right) \nu_{\mathfrak{Z}!} \sum_{\perp} (1 - \nu u) \nu_{\mathcal{C}} = {}^{17}\lambda \\ & \left(\nu_{\underline{A}} - \iota_{\mathcal{A}} \right) \left({}^{11}\zeta_{\underline{Z}} - \iota_{\mathcal{Z}} x \right) \nu_{\mathfrak{Z}!} \sum_{\perp} (1 - \nu u) \nu_{\mathcal{C}} = {}^{13}\lambda \\ & \left(\nu_{\underline{X}} - \iota_{\mathcal{X}} \right) \left({}^{11}\zeta_{\underline{Z}} - \iota_{\mathcal{Z}} x \right) \nu_{\mathfrak{Z}!} \sum_{\perp} (1 - \nu u) \nu_{\mathcal{C}} = {}^{15}\lambda \\ & \qquad \qquad \qquad \nu_{\mathcal{Z}}^{\iota_{\mathcal{X}}} \nu_{\mathcal{C}} = {}^{22}\lambda \\ & \qquad \qquad \qquad \nu_{\mathcal{Z}}^{\iota_{\mathcal{X}}} \nu_{\mathcal{C}} = {}^{12}\lambda \\ & \qquad \qquad \qquad \nu_{\mathcal{Z}}^{\iota_{\mathcal{X}}} \nu_{\mathcal{C}} = {}^{11}\lambda \\ & \qquad \qquad \qquad \text{avec } [{}^{\delta}\nu] = \dot{\nu} \sum \text{ et} \\ & \qquad \qquad \qquad \iota [(\nu u - 1) \nu q \nu_{\mathcal{Z}} - \nu_{\mathcal{Z}}] \end{aligned}$$
$$\xi = [x^{\nu_1} \bar{y}^{\nu_2} \bar{y}^{\nu_3} \dots \bar{y}^{\nu_{l-1}} \bar{y}^{\nu_l} x^{\nu_{l+1}} \dots x^{\nu_{l+m}}] = F_*(\xi).$$

$$\sum_{i \in A} u_i = \sum_{j=1}^n x_j$$

$$\cdot \left\{ \left({}^V \underline{x} - g(\underline{x}) \right) + \left({}^g \underline{x} - g(\underline{x}) \right) + \frac{({}^V f - 1)}{({}^V \underline{x} \cdot {}^g \underline{x})} \right\} \cdot {}^V q \left({}^V M - 1 \right) +$$

Dans cette annexe nous définissons $v^{\text{BL}}(\underline{y}^{\text{tr}})$. Sous le bootstrapap avec moyenne ajustée,

Annexe B

où $\hat{S}^{yx\prime\prime} = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})$. De même, $\text{Cov}(\bar{x}, \bar{y}) = -S_{xy}^{\prime\prime}/N$. Ceci complète la preuve de la convergence de $\sum_{i=1}^n \frac{y_i}{x_i} \frac{x_i}{y_i}$ pour $\sum_{i=1}^n \frac{y_i}{x_i} \frac{x_i}{y_i}$.

$$N/N_{\text{ex}} =$$

$$E_{\cdot}^{\cdot}(\underline{y})\{f(\underline{x}) + (f(\underline{x}) - 1)/(\underline{x} - \underline{x})^{\cdot} f(\underline{x}) + \underline{x}\}^{\cdot} \underline{y} = \underline{y}^{\cdot} \underline{f} - ((f(\underline{x}) - 1)/(\underline{x} - \underline{x})^{\cdot} f(\underline{x}) + \underline{x})^{\cdot} \underline{y}^{\cdot} \underline{f}$$

$$E_{\gamma^*, \gamma}(\bar{y}, E_{\gamma^*, \gamma|B}(\bar{x}_B)) - E_{\gamma^*, \gamma}(x_B, y) =$$

$$Cov(x^B, y^A) = E(y^A \cdot x^B) - E(y^A) \cdot E(x^B)$$

réalisé indépendamment. Un échantillon est divisé en classes d'imputation C_l ($l = 1, \dots, L$) dans chacune desquelles on suppose que le taux de réponse est uniforme et on procède à l'imputation. Une classe d'imputation peut recouper les strates. Nous supposons aussi que la classe d'imputation à laquelle appartient une unité échantillonnée est identifiée correctement avant l'imputation. Dénotons les nombres d'unités échantillonnées et de répondants dans $S_h \cap C_l$ comme étant n_{hl} et r_{hl} , respectivement. Alors, on voit que, sachant n_{hl} et r_{hl} , le plan correspondant dans $S_h \cap C_l$ est le même que celui discuté dans le présent article si nous considérons les n_{hl} unités et les r_{hl} répondants comme étant $A + B$ et A , respectivement. Par conséquent, le bootstrap avec moyenne ajustée peut être exécuté indépendamment dans différents $S_h \cap C_l$ ($h = 1, \dots, H; l = 1, \dots, L$). La taille de $S_h \cap C_l$ dénotée par N_{hl} peut être estimée par $\hat{N}_{hl} = N_h(n_{hl}/n_h)$. Notons qu'il s'agit d'une méthode bootstrap conditionnée sur le nombre de répondants.

6. Conclusion

Dans le présent article, nous avons proposé le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La méthode requiert un simple ajustement de la moyenne et permet de traiter l'estimation des fonctions de répartition et de quantiles, car elle ne nécessite pas de rééchantillonnage. Le développement en série de Taylor montre que la méthode a de bonnes propriétés conditionnelles pour les estimateurs par le ratio et par la régression. Une étude par simulation démontre qu'elle a aussi des propriétés conditionnelles similaires lors de l'estimation des fonctions de répartition et des quantiles. Une extension à l'échantillonnage à deux phases stratifiées est simple. Conditionnellement aux tailles d'échantillon de première phase, la méthode permet de traiter l'échantillonnage à deux phases stratifié et l'imputation sous le mécanisme de réponse uniforme. Nous posons à l'heure actuelle une extension de la méthode proposée à des plans d'échantillonnage multiphasés plus généralisés.

Remerciements

Cette étude a été financée par une bourse de la Société japonaise de promotion de la science. L'auteur remercie le professeur Randy R. Sitter, le rédacteur en chef, le rédacteur adjoint et deux examinateurs de leurs commentaires et suggestions utiles.

Annexe A

Dans la présente annexe, nous montrons que la méthode bootstrap proposée fournit des estimations de la variance convergentes pour une classe d'estimateurs considérés par Rao et Sitter (1997). Nous utilisons les mêmes conditions que dans Rao et Sitter (1997) avec une notation légèrement différente. Pour simplifier, nous supposons qu'il n'existe qu'une seule strate, mais une extension à l'échantillonnage à deux phases stratifié est simple.

Considérons une classe d'estimateurs, $\theta = h(\bar{y}^A, \bar{x}^A, \bar{x}^B)$, d'un paramètre de population $\theta = h(\bar{Y}, \bar{X}, \bar{X})$, où \bar{Y} et \bar{X} sont les moyennes de population des vecteurs y et x , c'est-à-dire $\bar{Y} = N^{-1} \sum_{i \in P} y_i$, et $\bar{X} = N^{-1} \sum_{i \in P} x_i$. Ici, x est observé dans l'échantillon de première phase $A + B$, tandis que y est mesuré uniquement dans l'échantillon de deuxième phase A . Les moyennes d'échantillon (\bar{y}^A, \bar{x}^A) et \bar{x}^B sont calculées dans A et B , respectivement, c'est-à-dire $\bar{y}^A = n_A^{-1} \sum_{i \in A} y_i$, $\bar{x}^A = n_A^{-1} \sum_{i \in A} x_i$, et $\bar{x}^B = n_B^{-1} \sum_{i \in B} x_i$.

Par un développement en série de Taylor, nous obtenons

$$\theta = \theta + \nabla h(\Delta \bar{y}^A, \Delta \bar{x}^A, \Delta \bar{x}^B) + o_p(n^{1/2}),$$

où ∇h est le vecteur de gradients de h évalué à $(\bar{Y}, \bar{X}, \bar{X})$, $\Delta \bar{y}^A = \bar{y}^A - \bar{Y}$, $\Delta \bar{x}^A = \bar{x}^A - \bar{X}$, et $\Delta \bar{x}^B = \bar{x}^B - \bar{X}$, 33.7 de Rao et Sitter 1997, page 757 et les conditions requises). Alors, la variance de $\theta = h(\bar{y}^A, \bar{x}^A, \bar{x}^B)$ est approximée par

$$V(\theta) \approx \Delta h' \sum_{(y^A, x^A, x^B)} \Delta h,$$

où $\sum_{(y^A, x^A, x^B)} \Delta h$ est la matrice de variance-covariance de $(\bar{y}^A, \bar{x}^A, \bar{x}^B)$ sous échantillonnage à deux phases répété. Comme A et B sont des EASSR de taille n_A et n_B tirés de la population P , respectivement, nous voyons que $\sum_{(y^A, x^A, x^B)} S_y^2 = (1 - f_A) S_y^2 / n_A$ et $\sum_{(y^A, x^A, x^B)} S_x^2 = (1 - f_B) S_x^2 / n_B$, où $S_y^2 = (N - 1)^{-1} \sum_{i \in P} (y_i - \bar{Y})^2$ et $S_x^2 = (N - 1)^{-1} \sum_{i \in P} (x_i - \bar{X})^2$. Pour la population A à partir de P et pour le choix de sélection d'un EASSR A à partir de P et pour le choix d'un EASSR B à partir de $P - A$ sachant A , respectivement. Notons que $E_{B|A}(x_B) = (\bar{X} - f_A \bar{X}) / (1 - f_A)$.

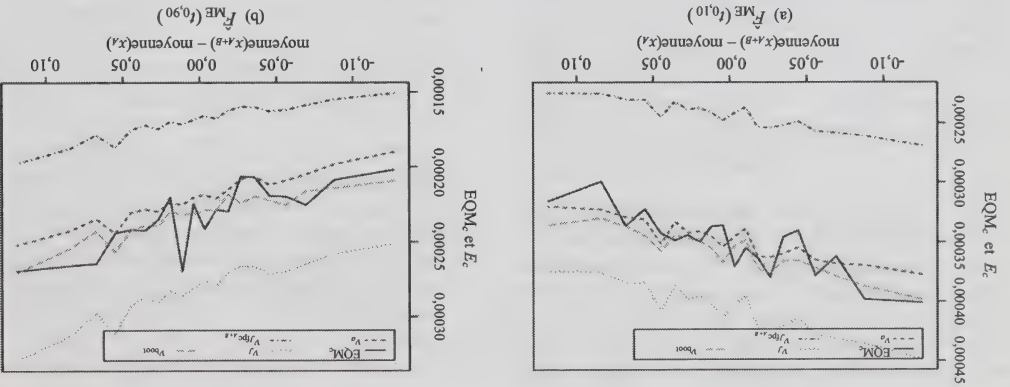
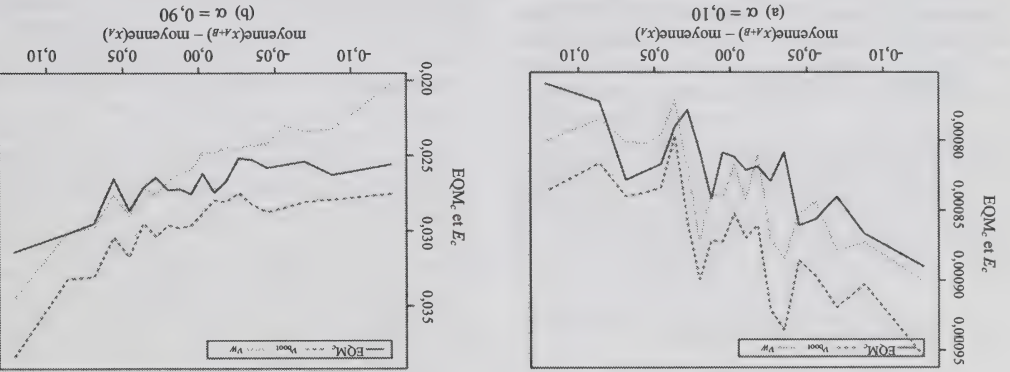
Donc, nous avons

$$\text{Cov}(\bar{y}^A, \bar{x}^B) = E(\bar{y}^A \bar{x}^B) - E(\bar{y}^A) E(\bar{x}^B)$$

$$= E(\bar{y}^A E_{B|A}(x_B)) - \bar{Y} \bar{X}$$

$$= -S_{yx}^2 / N,$$

où $S_{yx}^2 = (N - 1)^{-1} \sum_{i \in P} (y_i - \bar{Y})(x_i - \bar{X})$. De même, $\text{Cov}(\bar{x}^A, \bar{x}^B) = -S_x^2 / N$.

Figure 1 EQM_c et $E_c(v)$ pour $F_{ME}(t_{\alpha})$ Figure 2 EQM_c et $E_c(v)$ pour l'estimation des quantiles

5. Remarques supplémentaires

5.1 Échantillonnage à deux phases stratifié

Supposons qu'une population doit être stratifiée en H strates, mais qu'on ne dispose d'aucune information pour la stratification. Une solution possible dans cette situation est de commencer par obtenir un EASSR de taille n' à partir de la population, d'observer les variables auxiliaires, y compris celles pour la stratification, de stratifier l'échantillon en H strates et, dans chaque strate, de tirer un EASSR de taille n_h à partir de n'_h unités appartenant à la strate h dans l'échantillon. Voir, par exemple, Cochran (1977, section 12.2) pour les détails.

Soit N'_h la taille de la strate h dans la population. Sous la condition $n'_h > 0$, l'échantillonnage de première phase dans la strate h décrit plus haut est équivalent à l'échantillonnage aléatoire simple sans remise de taille n'_h dans la strate h réalisé de façon indépendante dans chacune des strates.

5.2 Non-réponse

Le commentaire qui précède s'applique aux données

d'enquête imputées sous le mécanisme de réponse unit-forme. Supposons qu'une population est stratifiée en S_h ($h = 1, \dots, H$) où l'échantillonnage aléatoire simple est

La figure 2 montre les propriétés conditionnelles de $v^{boot}_{ME}(\alpha)$ et de $v^{boot}_{ME}(\alpha)$ pour $\alpha = 0.10, 0.90$. Nous voyons que $v^{boot}_{ME}(\alpha)$ et $v^{boot}_{ME}(\alpha)$ suivent tous deux l'EQM, de la même façon, quoique le premier possède uniformément un biais par excès.

Estimateur		α	
0.10	0.25	0.50	0.75
%BiAs	6.27	14.32	10.05
CV	0.53	0.51	0.52
$V_M(f_{ME}^{-1}(\alpha))$	1.64	3.75	2.92
CV	0.50	0.45	0.46
			0.52

Tableau 2 Estimation de la variance pour les quantiles

Le tableau 2 résume les résultats pour l'estimation des quantiles. Il démontre que le bootstrap avec moyenne ajustée produit un biais par excès dans l'estimation de $V(f_{-1}^{ME}(\alpha))$, mais un biais négligeable dans l'estimateur de la variance de Woodruff.

$$\left[\frac{F_{-1}^{ME}(\alpha + \zeta_{1-K/2} \hat{\phi}^f) - F_{-1}^{ME}(\alpha - \zeta_{1-K/2} \hat{\phi}^f)}{2\zeta_{1-K/2}} \right] = \left[\frac{F_{-1}^{ME}(\alpha) - F_{-1}^{ME}(\alpha - \zeta_{1-K/2} \hat{\phi}^f)}{2\zeta_{1-K/2}} \right],$$

Par inversion directe de $F_{ME}^{\text{ME}}(t)$, nous estimons le quantile α . Pour obtenir \hat{p} , pour $F_{ME}^{\text{ME}}(t)$, nous fixons t à la valeur \hat{t}_α , où $\hat{t}_\alpha = \inf \{t : n^{-1} \sum_{i=1}^n I(Y_i \leq t) \geq \alpha\}$, un estimateur utilisant uniquement $\{Y_i : i \in A\}$. Pour l'estimation de la variance, nous avons créé $K = 1\,000$ échantillons bootstrap. En vue de comparaison, nous avons également calculé l'estimateur de la variance de Woodruff (Woodruff 1952 et Shao et Tu 1995, page 238),

4.3 Estimation des quantiles

figure 1 montre la représentation graphique d'EOM_c en fonction des moyennes de groupe de $X^{A+B} - X^A$ pour t et $t^{90\circ}$. On constate que $V(f^{ME}(t))$ et $V(f^{ME}(t^{90\circ}))$ ont tous deux le même comportement conditionnellement à $X^{A+B} - X^A$. Les estimateurs jackknife de la variance, $V_f(f^{ME}(t))$ et $V_f^{diff}(f^{ME}(t))$, quoique biaisés, suivent une tendance de l'EOM_c.

En nous inspirant de Royall et Cumberland (1981a, 1981b), nous avons ordonné les $M = 5\,000$ échantillons simulés sous les valeurs de $x^{-\nu+B}$ en vingt groupes consécutifs de $G = 250$ dans chacun desquels l'EQM conditionnelle (EQM_c) simulée et la moyenne conditionnelle de $v(E_c(v))$ ont été calculées. La

Estimateur	0.10	0.25	0.50	0.75	0.90
$\nu_{\text{boot}}(f^{ME}(a))$ %Biais	0.27	-0.22	0.64	0.83	2.73
CV	0.19	0.14	0.15	0.24	
$\nu_a(f^{ME}(a))$ %Biais	-2.29	-2.03	-0.47	-1.95	-3.26
CV	0.17	0.11	0.09	0.19	0.19
$\nu_f(f^{ME}(a))$ %Biais	14.24	17.29	22.98	23.80	24.97
CV	0.24	0.21	0.25	0.27	0.36
$\nu_{f^{jpc}}(f^{ME}(a))$ %Biais	-31.45	-29.63	-26.21	-25.72	-25.02
CV	0.33	0.30	0.27	0.27	0.30

Tableau 1 Estimation de la variance pour l'EMV pseudo-

Le tableau I présente le biais relatif (%biais) et le coefficient de variation (CV) des quatre estimateurs de la variance pour $F_{ME}^2(r_0)$ ($\alpha = 0,10, 0,25, 0,50, 0,75, 0,90$), où $F^2(r_0) = \alpha$. Ici, %Biais et CV ont été calculés sous la forme %Biais = $100 \times (M^{-1} \sum_{m=1}^M v_m - E(M)/E(M))$ et $CV = [M^{-1} \sum_{m=1}^M (v_m - E(M))^2]^{1/2} / E(M)$, respectivement, où v_m est une estimation de la variance dans la m^{e} exécution de la simulation. Le tableau I démontre que $F_{ME}^2(r_0)$ présente un biais par excès, puis que les fractions d'échantillonnage ne sont pas négligeables, que d'ajustement *ad hoc* $(1 - F^{+B})$ est trop faible, et que $F_{ME}^2(r_0)$ et $F_{ME}^2(r_0)$ sont tous deux approchés mativement sans biais, quoique le dernier soit un peu plus instable, ce qui est typique d'une méthode de rééchan-

tion pour population finie *ad hoc* est $v_{fjpc}(F^{ME}(t))(1 - f^{A+B} v_j^F(F^{ME}(t)))$.

ou $\theta = F_{ME}(t)$, $(\theta^{(-)})_{t \in \mathcal{T}}$ est la pseudo-estimation par jackknife et $\theta^{(1)} = n^{-1} \sum_{i=1}^n \Delta_{i+48}^{(-)}$. Notons que, pour $j \in \mathcal{A}$, y_j et x_j sont toutes deux éliminées de l'échantillon, tandis que pour $j \in \mathcal{B}$, seul x_j est éliminé (voir Rao et Sitter 1995 et Sitter 1997). La formule avec correc-

$$\sum_{j \in A+B} \frac{n^{A+B}}{(1-\theta)^{A+B}} = (\theta)^{A+B} (1-\theta)^{A+B} = 1$$

sous échantillonnage à deux phases proposé par Wu et Luan (2003) et défini par

$$F_{ME}^{(t)}(t) = \sum_{i \in A} p_i I(Y_i \leq t), \quad (7)$$

où p_i maximise la fonction de pseudo-vraisemblance $l(p) = \sum_{i \in A} (N/n_i) \log p_i$ sous les contraintes a) $\sum_{i \in A} p_i = 1$ ($0 < p_i < 1$); et b) $\sum_{i \in A} p_i g_i = n_{A+B}^{-1} \sum_{i \in A+B} g_i$ ou $g_i(x_i, t) = P(Y \leq t | x_i)$ sous un certain modèle de travail. Par exemple, nous pouvons supposer que $\log(g_i/(1 - g_i)) = x_i' \theta$ avec une fonction de variance $A(\theta) = g(1 - g)$. Chen, Sitter et Wu (2002) ont montré un algorithme simple pour le calcul de p_i . Il peut être démontré (voir Wu et Luan 2003) que, sous l'échantillonnage à deux phases considéré dans le présent article,

$$F_{ME}^{(t)}(t) = n_{A+B}^{-1} \sum_{i \in A} I(Y_i \leq t) + \left\{ \beta + o_p(n_{A+B}^{-1/2}) \right\} + \sum_{i \in A+B} g_i - n_{A+B}^{-1} \sum_{i \in A} g_i$$

où $\beta = \sum_{i \in A} (g_i - \bar{g}) I(Y_i \leq t) / \sum_{i \in A} (g_i - \bar{g})^2$ avec $\bar{g} = (N^{-1} \sum_{i \in A} g_i)$. Notons que cette équation n'est pas utilisée dans l'estimation, mais elle montre que la variance de $F_{ME}^{(t)}(t)$ peut être estimée par le bootstrap avec moyenne ajustée, puisque $F_{ME}^{(t)}(t)$ est approximé par un estimateur de type régression.

3.4 Estimation des quantiles

L'estimation des quantiles peut être obtenue directement en inversant $F^{(t)}(t)$ par $F^{-1}(\alpha) = \inf \{t : F^{(t)}(t) \geq \alpha\}$ pour un certain $\alpha \in (0, 1)$. Par exemple, si on utilise (7), alors une estimation des quantiles est donnée par $y^{(k)}$, où $y^{(k)}$ est la statistique de k -ième ordre de y telle que $\sum_{i=1}^{k-1} p_i^{(t)} < \alpha$ et $\sum_{i=1}^k p_i^{(t)} \geq \alpha$ (Chen et Wu 2002). Sous certaines conditions spécifiées dans Chen et Wu (2002), une représentation de type Bahadur de $F_{ME}^{(t)}(\alpha)$ peut être établie. Donc, l'estimateur de la variance par le bootstrap avec moyenne ajustée pour $F_{ME}^{(t)}(\alpha)$ est convergent par rapport au plan. Notons qu'il n'existe aucune forme explicite de l'estimateur de la variance pour $F_{ME}^{(t)}(\alpha)$, mais qu'on peut appliquer un estimateur de la variance convergent basé sur l'estimation d'intervalle de Woodruff (Woodruff 1952).

4. Simulation

4.1 Population et échantillonnage

Nous avons réalisé une étude par simulation afin d'examiner l'estimation de la variance par le bootstrap avec moyenne ajustée pour les estimateurs de la section 3. Nous présentons ici les résultats pour l'estimation des fonctions de répartition et des quantiles. Les résultats pour les estimateurs

4.2 Estimation des fonctions de répartition

par le ratio et par la régression peuvent être obtenus auprès de l'auteur sur demande.

Pour commencer, nous avons généré la variable aléatoire x pour une population finie P de taille $N = 2\,000$ en utilisant une loi Gamma(1, 1). La variable dépendante y a ensuite été générée au moyen de $y_i = x_i + \sqrt{x_i} v_i$, où $v_i \sim N(0, 0.5^2)$. Un EASSR $A+B$ de taille $n_{A+B} = 800$ a été sélectionné à partir de la population, puis un EASSR A de taille $n_A = 200$ a été sélectionné à partir de $A+B$. La population est demeurée fixe au cours des exécutions de la simulation, puisque nous nous concentrons sur les propriétés de l'échantillonnage répété par rapport au plan.

Pour l'estimation des fonctions de répartition, nous avons pris $F_{ME}^{(t)}(t)$ comme exemple. D'autres estimateurs, comme ceux de Chambers et Dunstan (1986) et de Rao, Kovari et Mantel (1990) peuvent être traités de la même façon quand un estimateur approximativement sans biais par rapport au plan. Nous avons supposé que le modèle de travail pour g dans $F_{ME}^{(t)}(t)$ était le logit avec variance binomiale. L'estimateur bootstrap de la variance $v_{boot}(F_{ME}^{(t)}(t))$ a été calculé avec $K = 200$. Nous avons utilisé le BBE pour construire un échantillon bootstrap. Le nombre total de simulations était $M = 5\,000$, tandis que l'EQM réelle de $F_{ME}^{(t)}(t)$ à un temps t donné a été estimée sur 50 000 exécutions.

Nous avons comparé $v_{boot}(F_{ME}^{(t)}(t))$ à trois estimateurs de la variance : l'estimateur analytique de Wu et Luan (2003), le jackknife avec suppression d'une unité standard et le jackknife avec suppression d'une unité et une correction pour population finie *ad hoc*. L'estimateur de Wu et Luan (2003) est

$$v_D(F_{ME}^{(t)}(t)) = (n_{A+B}^{-1} - N^{-1}) S_2^t + (n_{A+B}^{-1} - n_{A+B}^{-1}) S_2^{D^t}$$

où les deux composantes S_2^t sont estimées respectivement par

$$S_2^t = s_2^2 + \left[\frac{1}{\sum_{j>t: i \in A+B} n_{ij}} \right] \sum_{j>t: i \in A+B} \frac{1}{n_{ij}} \sum_{j>t: i \in A} n_{ij} \left[\beta^{F^t} \right]$$

où $s_2^2 = \{n_A(n_A - 1)\}^{-1} \sum_{i < j: i \in A} v_{ij}^2$ et $\beta^{F^t} = \sum_{i < j: i \in A} v_{ij}^2 / \sum_{i < j: i \in A} (I_i - I_j)^2$ avec $I_i = I(Y_i \leq t)$ et $\bar{g}_i = \bar{g}(x_i, t)$ et $n_{ij} = (\bar{g}_i - \bar{g}_j)^2$ avec $D_i^t = (D_i - \bar{D})^2$ et $n_{ij} = \bar{g}_i(1 - \bar{g}_i) + \bar{g}_j(1 - \bar{g}_j)$ avec $D_i^t = I_i - \bar{g}_i$, $\beta = \sum_{i < j: i \in A} I_i(I_i - \bar{g}_j) / \sum_{i < j: i \in A} (\bar{g}_i - \bar{g}_j)^2$ et $\bar{g}_i = I_i - \bar{g}_i$. La formule du jackknife avec suppression d'une unité standard est donnée par

moynes d'échantillon et non des moyennes de population (voir le commentaire fait par Derrati et Rao 2004, page 21).

3.2 Estimateur par la régression

3.2 Estimateur par la régression

régression. L'estimateur de la moyenne de population est

$$\bar{y}_r = \frac{\sum y_r}{n} + b \left(\frac{\sum x_r}{n} - \bar{x} \right) = \frac{\sum y_r}{n} + b \left(\bar{x} - \bar{x}' \right), \text{ où}$$
$$b = \frac{s_{xy}/S^2_x}{s_y/s^2_x} = \frac{(n-1)^{-1} \sum y_r(x_r - \bar{x}')}{\sum (x_r - \bar{x}')^2} = w_r b^r (\bar{x} - \bar{x}').$$

En utilisant $\bar{y}_r = \frac{\sum y_r}{n} + (1-w_r)b^r(\bar{x}-\bar{x}')$. Soit $\bar{y}_r = \frac{\sum y_r}{n} + (1-w_r)b^r(\bar{x}-\bar{x}')$. Les résultats de l'annexe A (voir aussi l'annexe B), nous

[illegible]

de la variance de $\hat{Y}^{(j)}$ (Sitter 1997, page 781) est

$$(9) \quad \frac{(\iota - 1)}{\iota \partial^{(g+y)X}(\nu X - \iota X)} \sum \frac{\nu^{\exists!} (1 - g+y)u}{\nu q^y z} u + \frac{(\iota - 1)}{\iota \partial^{(\nu X - \iota X)}} \sum \frac{\nu u}{\nu z} + \frac{(\iota - 1)}{\iota \partial^{(\nu X - \iota X)}} \sum \frac{\nu u}{\nu z} + \frac{g+y}{\iota} \sum \frac{\nu u}{\nu q^y (g+y f - 1)} + \frac{\nu u}{\iota \partial^{(\nu f - 1)}} = (\iota \underline{f}) \Pi_A$$

ou $c_i = n_i + (x_i - \bar{x}) / (n_i - 1) S_{x_i}^2$, les valeurs d'effectif. Parant de (5) et (6), $v_i^{\text{book}}(y_i)$, $v_i^{\text{BL}}(y_i)$ et $v_i^{\text{LR}}(y_i)$ donnent des résultats similaires à condition que tous les c_i soient presque nuls et que le dernier terme ne soit pas négligable.

3.3 Estimation des fonctions de répartition

A titre d'exemple, prenons l'estimateur du maximum de vraisemblance pseudo-empirique calé sur un modèle (ME)

Nous montrons à l'annexe A que la méthode bootstrap proposée produit une estimation de la variance convergente par rapport au plan pour la classe d'estimateurs étudiés par Rao et Sitter (1997). Puisqu'aucun rééchantillonnement n'est effectué, la méthode s'applique aussi à l'estimation des fonctions de répartition. Sous certaines conditions de régularité pour la fonction de répartition de population, elle produit des estimateurs de la variance convergent par rapport au plan pour les quantiles.

3. Illustrations

L'estimateur par le ratio $\hat{y}_r = r \cdot \bar{x}^{y+rB}$, où $r = m^y / m^{y+B}$, et $\bar{x} = (1 + (1 - w^y) \bar{x}^B) / (1 + (1 - w^y) \bar{x}^B)$. Soit $\hat{y}_r = \bar{y}_r / \bar{x}^B$. En utilisant les résultats de l'annexe A avec $h(\hat{y}_r)$, $\bar{x}^B / \bar{x}^B = (1 + (1 - w^y) \bar{x}^B) / (1 + (1 - w^y) \bar{x}^B)$, nous pouvons approximer la variance de \hat{y}_r sous la méthode bootstrap proposée $V(\hat{y}_r)$ par

$$\begin{aligned}
(\varepsilon) \left[\frac{{}^{\mathcal{G}\mathcal{X}}\mathcal{S}}{\mathcal{Z}\mathcal{S}} \frac{({}^{\mathcal{V}}f-1)}{({}^{\mathcal{V}}\mathcal{M}-1)} + {}^{\mathcal{V}\mathcal{X}}\mathcal{S} \frac{({}^{\mathcal{V}}f-1)}{({}^{\mathcal{V}}f-{}^{\mathcal{V}}\mathcal{M})} \right] \mathcal{Z}\mathcal{A}^{\frac{{}^{\mathcal{G}+\mathcal{V}}\mathcal{U}}{({}^{\mathcal{G}+\mathcal{V}}f-1)} +} \\
{}^{\mathcal{V}\mathcal{X}\mathcal{P}}\mathcal{S}^{\mathcal{V}\mathcal{A}} \frac{{}^{\mathcal{G}+\mathcal{V}}\mathcal{U}}{({}^{\mathcal{G}+\mathcal{V}}f-1)} \frac{({}^{\mathcal{V}}\mathcal{X}/{}^{\mathcal{G}+\mathcal{V}}\mathcal{X})}{({}^{\mathcal{V}}\mathcal{X}/({}^{\mathcal{V}}f-1))} \mathcal{Z} + \\
\frac{{}^{\mathcal{V}}\mathcal{U}}{\mathcal{Z}\mathcal{S}} \frac{{}^{\mathcal{V}}\mathcal{A}}{({}^{\mathcal{V}}f-1)} \mathcal{Z} \frac{({}^{\mathcal{V}}\mathcal{X}/{}^{\mathcal{G}+\mathcal{V}}\mathcal{X})}{({}^{\mathcal{V}}\mathcal{X}/({}^{\mathcal{V}}f-1))} = \frac{({}^{\mathcal{V}}\mathcal{A})}{({}^{\mathcal{V}}\mathcal{X})} \mathcal{A}
\end{aligned}$$

où

$$S_2^{\mathcal{P}} = (n-1)^{-1} \sum_{i \in \mathcal{I}} (y_i^{\mathcal{P}} - \bar{y}^{\mathcal{P}})^2, \quad S_2^{\mathcal{V}} = (n-1)^{-1} \sum_{i \in \mathcal{I}} (y_i^{\mathcal{V}} - \bar{y}^{\mathcal{V}})^2$$

et

$$S_2^{\mathcal{B}} = (n-1)^{-1} \sum_{i \in \mathcal{B}} (x_i^{\mathcal{B}} - \bar{x}^{\mathcal{B}})^2, \quad S_2^{\mathcal{V}} = (n-1)^{-1} \sum_{i \in \mathcal{I}} (x_i^{\mathcal{V}} - \bar{x}^{\mathcal{V}})^2$$

de (3) peut être décrit comme un estimateur de la variance par « bootstrap-linéarisation ». Nous le dénotons par $v_{\mathcal{I}}(\underline{y}^{\mathcal{P}})$. Soulignons que $v_{\mathcal{I}}^{\text{BI}}(\underline{y}^{\mathcal{P}})$ est presque identique à l'estimateur jackknife-linéarisation de la variance de Rao et Sitter (1995),

$$(7) \quad \begin{aligned} & \frac{{}^{g+V}u}{\zeta^V \zeta^J} \frac{1}{({}^{g+V}f-1)} + \\ & {}^{Vg}S^V \frac{{}^{g+V}u}{({}^{g+V}f-1)} ({}^Vx / {}^{g+V}x) \zeta + \\ & \frac{{}^{Vg}S^V}{\zeta^V} \frac{1}{({}^Vf-1)} \zeta^V ({}^Vx / {}^{g+V}x) = ({}^Vx) \Gamma_A \end{aligned}$$

Puisqu'ils sont proches de $v_{1t}(\bar{x}_t)$, $v_{1t}(\bar{y}_t)$, son approximation de Monte Carlo $v^{\text{boot}}_t(\bar{x}_t)$ et $v^{\text{boot}}_t(\bar{y}_t)$ devraient donner de bons résultats non seulement inconditionnellement, mais conditionnellement à $(\bar{x}_t/\bar{y}_t, \bar{x}_t/\bar{y}_t)$ également. Il est intéressant de souligner que la linéarisation de Taylor dans la dérivation de $v^{\text{Bil}}_t(\bar{x}_t)$ est effectuée autour de

2. Bootstrap avec moyenne ajustée

Pour simplifier la notation, nous supposons ici qu'il n'existe qu'une seule strate. Pour étendre notre méthode à l'échantillonnage stratifié, il suffit de répéter la même procédure indépendamment dans diverses strates pour obtenir un échantillon bootstrap (voir Rao et Sitter 1997, pages 759 à 762).

Soit P l'ensemble d'étiquettes d'unité dans une population de taille N . Supposons que l'on sélectionne un échantillon aléatoire simple sans remise (EASSR) de taille n_{A+B} à partir de P et dénotons les étiquettes échantillonnées par $A+B$. La variable auxiliaire (le vecteur de variables auxiliaires) x_i est observée pour $i \in A+B$. Puis, nous tirons un EASSR de deuxième phase de taille $n_A < n_{A+B}$ à partir de $A+B$ et dénotons les étiquettes échantillonnées par A . La caractéristique (le vecteur de caractéristiques) y_i est mesurée pour $i \in A$. Soit $B = (A+B) - A$, $n_B = n_{A+B} - n_A$, $y_A = \{y_i : i \in A\}$, $x_A = \{x_i : i \in A\}$, et $x_B = \{x_i : i \in B\}$. Nous supposons qu'un estimateur approximativement sans biais par rapport au plan du paramètre θ peut s'écrire sous la forme $\hat{\theta} = t(y_A, x_A, x_B)$.

Sous la méthode proposée, nous construisons un échantillon bootstrap comme il suit.

1. Considérer A comme un EASSR de taille n_A tiré de P . Choisir n_A unités à partir de A par une méthode bootstrap appropriée pour un EASSR de taille n_A tiré de P . Dénoter les étiquettes échantillonnées par A^* .

2. Considérer B comme un EASSR de taille n_B tiré de $P - A$, sachant que A a été sélectionné. Choisir n_B unités à partir de B par une méthode bootstrap appropriée pour un EASSR de taille n_B tiré de $P - A$. Dénoter les étiquettes échantillonnées B^* .
3. Pour $j \in B^*$, définir l'ajustement de la moyenne comme étant \bar{x}_j , où

$$\bar{x}_j = x_j + f_A(\bar{x}_A - \bar{x}_A^*) / (1 - f_A), \quad (1)$$

$$\text{avec } \bar{x}_A = n_A^{-1} \sum_{i \in A} x_i, \bar{x}_A^* = n_A^{-1} \sum_{i \in A^*} x_i, \text{ et } f_A = n_A / N.$$

4. Soit $y_A^* = \{y_i : i \in A^*\}$, $x_A^* = \{x_i : i \in A^*\}$, et $\bar{x}_B^* = \{\bar{x}_j : j \in B^*\}$. L'analogue bootstrap de $\hat{\theta}$ est alors donné par $\hat{\theta}^* = t(y_A^*, x_A^*, \bar{x}_B^*)$.

Pour les méthodes bootstrap applicables à une population finie, voir Shao et Tu (1995, chapitre 6). Le bootstrap de Bernoulli (BBE) proposé par Fumakoka, Saigo, Sitter et Toida (2006) mentionnerons plus loin. Pour obtenir un échantillon bootstrap A^* dans le BBE, nous procédons à un

grand nombre K de fois et utiliser

$$v^{\text{boot}}(\theta) = K^{-1} \sum_{k=1}^K (\hat{\theta}^{(k)} - \hat{\theta}^{(1)})^2, \quad (2)$$

où $\hat{\theta}^{(k)}$ est la valeur de $\hat{\theta}^*$ dans le $k^{\text{ème}}$ échantillon bootstrap et $\hat{\theta}^{(1)} = K^{-1} \sum_{k=1}^K \hat{\theta}^{(k)}$.

Quand f_A est négligeable, l'ajustement de la moyenne (1) est inutile. La méthode susmentionnée se réduit alors pour un grand n_A à celle de Schreuder et coll. (1987) et de Sitter (1997).

La méthode bootstrap proposée est motivée par les deux plans d'échantillonnage I et II sont $[P \rightarrow A+B, A+B \rightarrow A]$ et $[P \rightarrow A, P-A \rightarrow B]$, respectivement, où \rightarrow signifie que « le deuxième membre est un EASSR provenant du premier membre ». Alors, I et II implémentent le plan de sondage identique. En fait, la probabilité d'échantillonnage attribué à un échantillon particulier $\{i_1, i_2, \dots, i_{n_A}\} \in A$, $j = (j_1, j_2, \dots, j_{n_B}) \in B$ dans I est $\Pr\{i \in A, j \in B\} = [N C_{n_A+B}^{n_A} C_{n_B}^{n_A}]/N!$, tandis qu'elle est $\Pr\{i \in A, j \in B\} = [N C_{n_A+B}^{n_A} C_{n_B}^{n_A}]/N!$ dans II. De toute évidence, la distribution d'échantillonnage d'un estimateur sous échantillonnage répété dépend du plan d'échantillonnage. Donc, il est commode de supposer que II est réalisé, même si I est employé.

En deuxième lieu, pour justifier l'ajustement de la moyenne (1), observons que la moyenne de x de l'échantillonnage répété sachant A , est $\bar{x}_{P-A} = (\bar{x} - f_A \bar{x}_A) / (1 - f_A)$. La valeur bootstrap de \bar{x}_{P-A} est donnée par $\bar{x}_{P-A}^* = (\bar{x} - f_A \bar{x}_A^*) / (1 - f_A)$. Donc, l'équation (1) équivaut à $\bar{x}_j = x_j - \bar{x}_{P-A} + \bar{x}_{P-A}^*$, un ajustement de la moyenne semblable à celui proposé par Rao et Shao (1992) dans le contexte de l'imputation hot deck sous mécanisme de réponse uniforme. Cet ajustement de la moyenne fait en sorte qu'existent les corrélations appropriées entre x dans A^* et x dans B^* nécessaires pour que l'estimation de la variance soit convergente lorsque les fractions d'échantillonnage sont élevées (voir Rao et Sitter 1997, page 760). Notons que la condition $n_A = n$, ou $f_A = f$, est essentielle à l'annulation de \bar{x} dans l'ajustement de la moyenne. Par conséquent, le bootstrap avec moyenne ajustée requiert une méthode bootstrap pour l'EASSR qui retient la taille originale d'échantillon, telle que le BBE.

Bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases

Hiroshi Saigo¹

Résumé

L'échantillonnage à deux phases est un plan utile lorsque l'on ne dispose pas de variables auxiliaires a priori. L'estimation de la variance sous ce plan est toutefois compliquée, particulièrement si les fractions d'échantillonnage sont grandes. Le présent article décrit une méthode bootstrap simple pour l'échantillonnage aléatoire simple à deux phases sans remise à chaque phase avec fraction d'échantillonnage élevée. Elle est applicable à l'estimation des fonctions de répartition et des quantiles, puisqu'aucune remise à l'échelle n'est effectuée. La méthode peut être étendue à l'échantillonnage à deux phases stratifié en répétant indépendamment la procédure proposée dans diverses strates. L'estimation de la variance de certains estimateurs classiques, comme les estimateurs par le ratio et par la régression, est étudiée à titre d'exemple. Une étude par simulation est réalisée pour comparer la méthode proposée aux estimateurs de la variance existants pour l'estimation des fonctions de répartition et des quantiles.

Mots clés : Échantillonnage double; rééchantillonnage; estimation de la variance.

1. Introduction

L'échantillonnage à deux phases ou échantillonnage double est un outil puissant pour l'estimation efficace dans les sondages. Habituellement, on tire un grand échantillon de première phase où les variables auxiliaires, corrélées aux caractéristiques d'intérêt et relativement faciles à obtenir, sont observées. Puis, on sélectionne un petit sous-échantillon à partir de l'échantillon de première phase pour mesurer les caractéristiques d'intérêt qui sont plus difficiles à obtenir. À l'étape de l'estimation, les variables auxiliaires de la première phase sont utilisées pour obtenir un estimateur efficace.

Une formule explicite de la variance d'échantillon d'un estimateur peut être compliquée, voire même inexistante sous échantillonnage à deux phases. Par conséquent, les méthodes de rééchantillonnage, comme le jackknife et le bootstrap, sont séduisantes dans ces conditions. Rao et Sitter (1995) et Sitter (1997) ont étudié l'approche du jackknife avec suppression d'une unité pour les estimateurs par le ratio et par la régression sous échantillonnage à deux phases et constaté que la méthode produit des estimations de la variance convergentes par rapport au plan ayant des propriétés conditionnelles désirables sachant les variables auxiliaires. Une faiblesse du jackknife avec suppression d'une unité est qu'il ne permet pas de traiter l'estimation des quantiles. De surcroît, l'intégration de la correction pour population finie dans l'estimation de la variance par le jackknife sous échantillonnage à deux phases n'est pas une question triviale (voir Lee et Kim 2002 et Berger et Rao 2006). Le bootstrap, par contre, élimine ces problèmes s'il est formulé convenablement.

Plusieurs méthodes bootstrap ont été proposées et étudiées pour l'échantillonnage à deux phases. Schreuder, Li et Scott (1987), Biemer et Atkinson (1993) et Sitter (1997) ont considéré des méthodes bootstrap similaires qui fournissent une estimation de la variance convergente lorsque les fractions d'échantillonnage sont négligeables. Rao et Sitter (1997) ont proposé un bootstrap avec rééchantillonnage pour les fractions d'échantillonnage élevées. Un inconvénient de l'approche de rééchantillonnage est qu'elle ne permet pas de traiter l'estimation des fonctions de répartition ni des quantiles. Dans le présent article, nous proposons un bootstrap corrigé sur la moyenne pour l'échantillonnage à deux phases qui permet l'estimation des fonctions de répartition et des quantiles. La méthode est simple et englobe les méthodes existantes pour les fractions d'échantillonnage négligeables à titre de cas particuliers. Récemment, Kim, Navarro et Fuller (2006) ont étudié l'estimation de la variance par rééchantillonnage sans rééchantillonnage pour l'échantillonnage à deux phases dans un cadre plus généralisé que celui du présent article. Toutefois, notre méthode diffère en ce que la correction pour population finie y est intrinsèque.

La présentation de l'article est la suivante. La section 2 décrit le bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases. La section 3 illustre le fonctionnement de la méthode proposée pour certains estimateurs classiques. La section 4 décrit l'exécution d'une simulation pour l'estimation des fonctions de répartition et des quantiles. La section 5 comprend la discussion d'autres applications du bootstrap avec moyenne ajustée. Enfin, les conclusions sont présentées à la section 6.

- Heldal, J. (1992). A method for calibration of weights in sample surveys. Dans *Workshop on uses of auxiliary information in surveys*. University of Örebro, Suède.
- Hidiroglou, M., et Patac, Z. (2004). Estimation par domaine par la régression linéaire. *Techniques d'enquête*, 30, 73-85.
- Lazafeld, P.F., et Menzel, H. (1961). On the relation between individual and collective properties. Dans *Complex Organizations: A Sociological Reader*. Holt, Reinhart and Winston. 422-440.
- Lemaitre, G., et Dufour, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- Lucy, D.M. (1986). Weighting sample survey data under linear constraints on the weights. Dans *Proceedings of the Social Statistics Section, American Statistical Association*, (Alexandria, VA), 325-330.
- Nieuwenbroek, N. (1993). *An integrated method for weighting characteristics of persons and households using the linear regression estimator*. Netherlands Central Bureau of Statistics.
- Särndal, C., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.
- Silva, P.L.N., et Skinner, C. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.
- Tam, S.M. (1995). Optimal and robust strategies for cluster sampling. *Journal of the American Statistical Association*, 90, 379-382.

Or, B_C est une régression par les moindres carrés ordinaires de r_{g1} sur x_{g1} , de sorte que

$$\sum_{g \in \mathcal{L}_1} (r_{g1} - B_C^T x_{g1}) x_{g1} = 0.$$

Donc, (14) devient

$$\text{var}_p[T_p] - \text{var}_p[T_H] =$$

$$\frac{M_2}{M} \left(1 - \frac{m}{M} \right) (M - 1)^{-1} B_C^T \sum_{g \in \mathcal{L}_1} x_{g1} x_{g1}^T B_C.$$

Preuve du théorème 3

L'estimateur GREG est invariant sous des transformations linéaires inversibles des variables auxiliaires. Donc, le modèle (9) peut être reparamétrisé pour donner

$$E_M[V_i] = \phi_1^T x_g + \phi_2^T (x_i - \bar{x}_g) \quad (15)$$

ou, de manière équivalente,

$$E_M[V_i] = \phi^T z_i$$

où

$$z_i = \begin{pmatrix} \bar{x}_g \\ x_i - \bar{x}_g \end{pmatrix}$$

et

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}.$$

Les paramètres du modèle (15) sont reliés à ceux du modèle (9) par $\phi_1 = \gamma_1 + \gamma_2$ et $\phi_2 = \gamma_2$.

Partant de la définition 1, en notant que

$$s = \bigcup_{g \in \mathcal{L}_1} U_g$$

pour le plan supposé, l'estimateur par la régression généralisée sous le modèle (15) est

$$\hat{T} = \hat{T}_\pi + \sum_{i \in s} \hat{\phi}_i^T z_i - \sum_{i \in s} \pi_i^{-1} \hat{\phi}_i^T z_i$$

$$= \hat{T}_\pi + \sum_{i \in \mathcal{L}_1} \sum_{g \in \mathcal{L}_1} \{ \hat{\phi}_1^T x_g + \hat{\phi}_2^T (x_i - \bar{x}_g) \}$$

$$- \sum_{g \in \mathcal{L}_1} \sum_{i \in \mathcal{L}_1} \pi_i^{-1} \{ \hat{\phi}_1^T x_g + \hat{\phi}_2^T (x_i - \bar{x}_g) \}. \quad (16)$$

Cependant, $\sum_{i \in \mathcal{L}_1} (x_i - \bar{x}_g) = 0$ pour chaque g . Donc (16) devient

Remarquons que (17) n'inclut pas l'estimateur de ϕ_2 . Les estimateurs par les moindres carrés

$$\hat{T}_\pi + \hat{\phi}_1^T (T^X - \hat{T}^X) = \hat{T}_\pi + \hat{\phi}_1^T (T^X - \hat{T}^X) \quad (17)$$

$$= \hat{T}_\pi + \hat{\phi}_1^T \sum_{g \in \mathcal{L}_1} \sum_{i \in \mathcal{L}_1} \pi_i^{-1} \bar{x}_{g1}$$

$$= \hat{T}_\pi + \hat{\phi}_1^T \sum_{g \in \mathcal{L}_1} \sum_{i \in \mathcal{L}_1} \pi_i^{-1} \sum_{g \in \mathcal{L}_1} \pi_i^{-1} \bar{x}_g$$

$$\hat{T} = \hat{T}_\pi + \sum_{g \in \mathcal{L}_1} \sum_{i \in \mathcal{L}_1} \pi_i^{-1} \hat{\phi}_1^T \bar{x}_g - \sum_{g \in \mathcal{L}_1} \sum_{i \in \mathcal{L}_1} \pi_i^{-1} \hat{\phi}_1^T \bar{x}_g$$

Bibliographie

Donc, $\hat{\phi}_1$ est une solution de (7). Par conséquent, l'estimateur GREG pour le modèle (9) est égal à \hat{T}_H à condition que $c_i = a_g N_g$.

Alexander, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

Cholent, P. (1984). L'ajustement des séries infra-annuelles aux répères annuels. *Techniques d'enquête*, 10, 39-53.

Clark, R.G., et Steel, D.G. (2002). The effect of using household as a sampling unit. *Revue Internationale de Statistique*, 70 (2), 289-314.

Tableau 4 REQMR relative médiane pour les estimateurs de domaine pour une taille d'échantillon $m = 1\ 000$

Variable	Domaines âge-sexe			Domaines régionaux		
	T_1	T_P	T_{H1}	T_1	T_P	T_{H1}
Occupé(e)	12,74	7,92	7,93	7,90	29,89	30,20
Occupée F	13,12	8,32	8,36	8,34	34,64	35,03
Revenu	13,25	8,43	8,49	8,47	28,04	28,12
Faible revenu	21,17	18,77	18,96	18,94	42,71	42,85
Heures travaillées	14,56	10,69	10,76	31,24	93,30	94,37
Parent seul	96,20	96,33	97,64	10,72	92,99	93,52
Arthrite	24,94	20,94	21,12	13,31	12,94	13,02
Fumeur(se)	32,10	29,25	29,39	29,37	12,32	12,27
PA élevée	27,01	23,80	23,97	23,95	15,83	15,31
Santé passable/mauvaise	39,64	37,73	38,05	38,08	22,38	22,30
Alcool	25,58	21,42	21,53	21,58	12,73	12,70

Tableau 5 REQMR relative médiane pour les estimateurs de domaine pour une taille d'échantillon $m = 10\ 000$

Variable	Domaines âge-sexe			Domaines régionaux		
	T_1	T_P	T_{H1}	T_1	T_P	T_{H1}
Occupé(e)	3,77	2,35	2,32	2,31	8,85	8,85
Occupée F	3,86	2,43	2,43	2,42	10,30	10,26
Revenu	3,91	2,53	2,51	2,51	8,24	8,23
Faible revenu	6,31	5,63	5,62	5,61	12,67	12,68
Heures travaillées	4,29	3,15	3,15	3,12	9,26	9,25
Parent seul	28,40	28,26	28,29	28,23	27,11	27,14
Arthrite	7,40	6,26	6,27	6,27	3,98	3,85
Fumeur(se)	9,53	8,58	8,57	8,57	3,69	3,67
PA élevée	8,07	7,02	7,01	7,01	4,66	4,48
Santé passable/mauvaise	11,69	11,02	11,02	11,01	6,75	6,69
Alcool	7,74	6,43	6,43	6,43	3,87	3,85

4. Discussion

L'estimateur GREG au niveau de la personne standard produit des pondérations inégales à l'intérieur des ménages. Les estimateurs GREG au niveau du ménage peuvent être utilisés pour obtenir des pondérations intégrées au niveau du ménage et de la personne, ce qui est avantageux pour les enquêtes recueillant de l'information sur des variables au niveau du ménage ainsi qu'au niveau de la personne. Nous avons démontré dans le présent article que l'avantage pratique de la pondération intégrée découlant de l'utilisation d'un estimateur GREG au niveau du ménage est associé à une perte faible, voire nulle. Pour les grands échantillons, l'estimateur GREG au niveau du ménage a une variance sous le plan plus faible que l'estimateur GREG au niveau de la personne. Pour les échantillons plus petits, l'utilisation de l'estimateur GREG au niveau du ménage produit, au plus, un faible accroissement de la variance pour certaines

variables, parce que cet estimateur est équivalent à l'utilisation d'un modèle de régression contenant un plus grand nombre de paramètres. Par conséquent, si les pondérations intégrées améliorent la cohérence des données de sortie d'une enquête-ménage, l'adoption de l'estimateur GREG au niveau du ménage ne causera qu'un accroissement faible, voire nul, de la variance et du biais des estimateurs.

Remerciements

Les présents travaux ont été financés conjointement par l'Australian Research Council et l'Australian Bureau of Statistics. Les opinions exprimées ici ne reflètent pas forcément celles de ces organismes. Les auteurs remercient Julian England, Frank Yu et Ray Chambers de leurs commentaires constructifs.

Tableau 1 REQMR relative pour une taille d'échantillon de 1 000 ménages

Variable	REQMR en %					% d'amélioration de l'EQM				
	T_1	T_p	T_{H1}	T_{H2}	T_{H2}	T_{H1}	T_{H2}	T_{H1}	T_{H2}	T_{H2}
Occupé(e)	2,62	2,09	2,09	2,10	2,10	0,20 (0,26)	-0,28 (0,27)	2,62	2,09	2,09
Occupé F	3,78	3,05	3,01	3,02	2,63	0,33 (0,33)	2,09 (0,33)	4,24	3,53	4,13
Revenu	2,56	2,20	2,19	2,19	1,04	0,25 (0,24)	0,61 (0,19)	1,43	0,19	1,07
Faible revenu	5,04	4,87	4,89	4,90	-0,62	(0,20)	0,18 (0,15)	2,64	0,00	0,65
Heures travaillées	3,08	2,54	2,53	2,53	0,94	(0,28)	1,61 (0,24)	2,64	0,00	2,12
Parent seul	12,50	12,73	12,02	11,65	10,84	(0,62)	10,23 (0,57)	11,50	0,58	11,21
Arthrite	5,52	4,50	4,53	4,53	-1,38	(0,17)	-0,08 (0,09)	-0,13	0,07	0,08
Fumeur(se)	4,73	4,57	4,60	4,61	-1,64	(0,18)	-0,26 (0,08)	-0,06	0,07	0,16
PA élevée	6,80	5,30	5,35	5,36	-1,70	(0,17)	-0,31 (0,08)	-0,04	0,06	0,08
Santé passable/mauvaise	9,79	9,42	9,47	9,47	-1,16	(0,16)	-0,71 (0,12)	-0,05	0,08	0,10
Alcool	4,81	4,66	4,70	4,71	-1,77	(0,16)	-0,77 (0,12)	-0,31	0,08	0,14

Tableau 2 Amélioration de l'EQM de l'estimateur GREG au niveau du ménage T_{H1} comparativement à T_p

Variable % d'amélioration de l'EQM

$m = 500$ 1 000 2 000 5 000 10 000 ∞

Occupé(e) -0,65 (0,31) 0,20 (0,26) 1,02 (0,24) 0,90 (0,21) 2,17 (0,21) 1,85

Occupé F 1,22 (0,37) 2,63 (0,33) 2,59 (0,33) 3,53 (0,31) 4,24 (0,31) 4,13

Revenu -1,53 (0,31) 1,04 (0,25) 0,48 (0,24) 0,61 (0,19) 1,43 (0,19) 1,07

Faible revenu -2,45 (0,27) -0,62 (0,20) 0,02 (0,18) 0,18 (0,15) 0,00 (0,00) 0,65

Heures travaillées -0,26 (0,34) 0,94 (0,28) 1,72 (0,27) 1,61 (0,24) 2,64 (0,24) 2,12

Parent seul 7,81 (0,69) 10,84 (0,62) 10,74 (0,61) 10,23 (0,57) 11,50 (0,58) 11,21

Arthrite -3,01 (0,24) -1,38 (0,17) -0,34 (0,12) -0,08 (0,09) -0,13 (0,07) 0,08

Fumeur(se) -3,91 (0,25) -1,64 (0,18) -1,02 (0,12) -0,26 (0,08) -0,06 (0,07) 0,16

PA élevée -2,93 (0,24) -1,70 (0,17) -0,86 (0,12) -0,31 (0,08) -0,04 (0,06) 0,08

Santé passable/mauvaise -3,67 (0,25) -1,16 (0,16) -0,71 (0,12) -0,05 (0,08) 0,03 (0,06) 0,10

Alcool -4,22 (0,23) -1,77 (0,16) -0,77 (0,12) -0,31 (0,08) -0,21 (0,07) 0,14

Variable	% d'amélioration de l'EQM					% d'amélioration de l'EQM				
	$m = 500$	1 000	2 000	5 000	10 000	∞	T_{H1}	T_{H2}	T_{H1}	T_{H2}
Occupé(e)	-1,85 (0,35)	-0,28 (0,27)	1,25 (0,25)	1,05 (0,21)	2,22 (0,21)	1,98	1,85	2,22	1,98	2,22
Occupé F	0,28 (0,39)	2,09 (0,33)	2,71 (0,33)	3,55 (0,29)	4,50 (0,30)	4,31	4,24	3,55	4,50	4,31
Revenu	0,75 (0,24)	0,71 (0,22)	0,90 (0,17)	1,30 (0,16)	1,30 (0,16)	1,37	1,43	1,30	1,30	1,37
Faible revenu	-3,15 (0,30)	-1,12 (0,22)	-0,15 (0,18)	0,06 (0,15)	0,00 (0,00)	0,94	-2,93	0,00	0,00	0,94
Heures travaillées	-1,51 (0,35)	0,70 (0,28)	1,98 (0,25)	1,79 (0,21)	2,57 (0,22)	2,26	-0,26	1,79	2,57	2,26
Parent seul	14,70 (0,53)	16,31 (0,49)	16,39 (0,47)	15,41 (0,44)	16,44 (0,44)	16,35	14,70	15,41	16,44	16,35
Arthrite	-3,31 (0,26)	-1,57 (0,18)	-0,05 (0,13)	-0,12 (0,09)	-0,10 (0,07)	0,16	-3,31	-0,12	-0,10	0,16
Fumeur(se)	-3,82 (0,28)	-1,81 (0,20)	-0,69 (0,14)	0,21 (0,11)	0,28 (0,10)	0,57	-3,82	0,21	0,28	0,57
PA élevée	-3,20 (0,26)	-2,06 (0,18)	-1,12 (0,13)	-0,40 (0,09)	-0,05 (0,07)	0,12	-3,20	-0,40	-0,05	0,12
Santé passable/mauvaise	-4,02 (0,28)	-1,07 (0,18)	-0,57 (0,13)	-0,09 (0,09)	0,00 (0,07)	0,15	-4,02	-0,09	0,00	0,15
Alcool	-5,00 (0,26)	-2,15 (0,18)	-0,82 (0,13)	-0,49 (0,09)	-0,29 (0,08)	0,18	-5,00	-0,49	-0,29	0,18

Steel et Clark : Estimation par la régression au niveau de la personne et au niveau du ménage

qu'il est égal à $N/n \sum_{g \in \mathcal{G}} \sum_{i \in U_g} y_i$ pour l'échantillonnage en grappes avec échantillonnage aléatoire simple des ménages où n est la taille réalisée de l'échantillon de personnes.

Les variables comprennent la situation d'activité, l'état de santé et d'autres caractéristiques. Toutes les variables sont dichotomiques, sauf celle du revenu (revenu annuel en dollars australiens, basé sur les données déclarées pour les fourchettes de revenu du recensement). « Occupée (F) » est une variable indicatrice dont la valeur est 1 si une personne est occupée et de sexe féminin, et 0 autrement. Les six premières variables sont tirées du recensement de la population et les cinq autres, de l'enquête sur la santé de la population.

3.2 Résultats

Le tableau 1 donne la racine de l'erreur quadratique moyenne relative (REQMR) de \hat{T}_p , \hat{T}^{H_1} et \hat{T}^{H_2} , pour une taille d'échantillon de 1 000 ménages. Les REQMR sont exprimés en pourcentage du total réel de population. Les biais n'ont pas été tabulés, parce qu'ils représentaient une composante négligeable de l'EQM dans tous les cas. Le pourcentage d'amélioration de l'EQM de \hat{T}^{H_1} et de \hat{T}^{H_2} relativement à \hat{T}_p est également présenté. Les chiffres entre crochets sont les erreurs-types de simulation de ces pourcentages d'amélioration.

Pour la taille d'échantillon susmentionnée, T^{H1} et T^{H2} donnent des résultats un peu moins bons que T_p pour les variables relatives à la santé et un peu meilleurs pour la plupart des autres variables. Nous observons le gain le plus important pour l'estimation du nombre de parents seuls; cette variance a été réduite de 10,8 % et de 16,3 %, respectivement, en utilisant les deux estimateurs GREG au niveau du ménage. Pour toutes les autres variables, l'amélioration est faible ou bien l'estimateur GREG au niveau du ménage donne d'un peu moins bons résultats que l'estimateur GREG au niveau de la personne. L'efficacité due à l'utilisation d'un estimateur GREG au niveau du ménage plutôt que T_p n'est jamais supérieure à 2,2 %.

Le tableau 2 montre le pourcentage d'amélioration de l'EOM résultant de l'utilisation de \hat{T}^{HI} plutôt que \hat{T}^p pour diverses tailles d'échantillon. Pour chaque chiffre, l'erreur-type de simulation est indiquée entre crochets. Le tableau 3 donne le pourcentage d'amélioration résultant de l'utilisation de \hat{T}^{HI} plutôt que \hat{T}^p . Sont également présentés les pourcentages d'amélioration asymptotiques ($m = \infty$) basés sur l'approximation en grand échantillon de la variance d'un estimateur GREG. Pour les deux estimateurs GREG au niveau du ménage, le pourcentage d'amélioration augmenté généralement avec la taille d'échantillon. Pour $m = 500$, les estimateurs GREG au niveau du ménage sont

générallement prises que l'estimateur GREG au niveau de la personne, quoique jamais de plus de 5 %. Pour $m = 10\,000$, nous observons une amélioration pour plus de la moitié des variables. Les améliorations les plus importantes sont celles constatées pour les estimations du nombre de parents seuls (11,5 %) et de femmes occupées (4,2 %); toutes les autres améliorations sont faibles. \hat{T}_{H1} et \hat{T}_{H2} n'ont jamais une variance dépassant de plus de 0,2 % celle de \hat{T}_p pour $m = 10\,000$. En général, \hat{T}_{H2} donne de meilleurs résultats que \hat{T}_{H1} pour les échantillons de plus grande taille, comme on s'y attendrait d'après le théorème 1, mais l'inverse est également vrai pour les petites tailles d'échantillon.

En pratique, les estimations des descripteurs de la population présentent souvent un intérêt que les totaux de population. Le tableau 4 montre les propriétés des divers estimateurs pour les domaines âgés-ccxc (12 catégories d'âge) et les domaines régionaux, pour une taille d'échantillon de 1 000 ménages. L'ensemble de données de recensement comportait 49 régions. L'ensemble de données de l'enquête sur la santé de la population ne contenait aucune variable de région semblable, de sorte que nous avons utilisé à la place le quintile socioéconomique du district de collecte (une unité géographique constituée d'environ 200 ménages contigus). Pour produire les estimateurs de domaine, nous avons calculé les pondérations à partir de chaque estimateur, puis pris la somme pondérée sur l'ensemble de l'échantillon dans le domaine. Cela équivaut à l'estimateur par le ratio pour le domaine décrit au cas 1, section 2.1 de Hidroglou et Patak (2004). Nous avons suivi cette méthode parce qu'elle est la plus utilisée en pratique, car elle permet d'estimer nous les totaux de domaine et de population à l'aide d'un seul domaine plus efficaces existent (Hidroglou et Patak 2004, cas 2 à 6).

Dans chaque cas, nous présentons la REQMR médiane sur les domaines. Le tableau montre que les différences entre les trois estimateurs GREG sont faibles. Pour les domaines âge-sexe, les estimateurs GREG au niveau du ménage donnent d'un peu de meilleurs résultats que l'estimateur GREG au niveau de la personne pour les variables de recensement, et d'un peu moins bons pour les variables de l'enquête sur la santé de la population. Pour les estimations régionales, les estimateurs GREG au niveau du ménage sont un peu moins bons dans tous les cas. Le tableau 5 montre que les propriétés de l'estimateur GREG au niveau du ménage sont fort semblables à celles de T_p pour une taille d'échantillon de 10 000 ménages. Il convient de souligner que les théorèmes 1 et 2 ne s'appliquent pas aux estimateurs de domaine que nous avons utilisés.

L'ajout de paramètres au modèle peut accroître la variance de l'estimateur GREG, quoique cet effet soit négligeable pour les grands échantillons. Il est possible que les effets contextuels n'aient que peu de pouvoir prédictif, voire aucun, pour certaines variables. Le cas échéant, on s'attendrait à ce que \hat{T}_p donne d'un peu moins bons résultats que \hat{T}_p pour les grands échantillons, et à peu près les mêmes résultats pour les grands échantillons.

Le modèles contextuel (9) contient tous les éléments de x_i et tous les éléments de \bar{x}_g . Une autre solution consisterait à utiliser uniquement les éléments de x_i et \bar{x}_g qui sont significatifs, ou qui produisent des améliorations de la variance estimée d'un estimateur GREG. Un estimateur GREG basé sur ce genre de modèle aurait probablement une variance plus faible que les estimateurs examinés dans le présent article, mais ne donnerait pas de pondérations intégrées, à moins d'utiliser les mêmes éléments de x_i et \bar{x}_g .

3. Étude empirique

3.1 Méthodologie

Nous avons entrepris une étude par simulation en vue de comparer les estimateurs GREG aux niveaux de la personne et du ménage, \hat{T}_p et \hat{T}_h , pour une gamme de variables d'enquête. Nous avons utilisé deux populations, constituées de 187 178 ménages sélectionnés aléatoirement à partir du recensement de la population de l'Australie de 2001 et de 210 132 ménages provenant de l'enquête nationale sur la santé de la population australienne de 1995. Tous les adultes et enfants faisant partie de ces ménages ont été inclus dans l'étude. La taille moyenne des ménages était de 2,5 environ. Nous avons tirés des échantillons en grappes à partir de ces populations, en sélectionnant d'abord des ménages par échantillonnage aléatoire simple sans remise, puis tous les membres des ménages échantillonnés. Nous avons simulés 10 000 ménages. Dans chaque cas, 5 000 échantillons ont été sélectionnés. Les variables auxiliaires x_i correspondaient à des variables indicatrices du sexe selon le groupe d'âge (12 catégories). (Ce choix de x_i signifie que l'estimation GREG équivaut à une poststratification.) Nous avons calculé l'estimateur GREG au niveau de la personne avec $c_i = 1(\hat{T}_p)$, l'estimateur GREG au niveau du ménage avec $a_g = N_g^{-1}(\hat{T}_h)$, et l'estimateur GREG au niveau du ménage avec $a_g = 1(\hat{T}_h)$. Nous avons également inclus l'estimateur de Hajek

$$\hat{T}_i = N \left(\sum_{i \in s} \pi_i^{-1} y_i \right) / \left(\sum_{i \in s} \pi_i^{-1} \right)$$

Le résultat montre que la réduction de la variance due à l'utilisation de \hat{T}_h (avec $a_g = 1$) plutôt que \hat{T}_p est une forme quadratique en B_C . Donc, l'importance de l'amélioration dépend de la mesure dans laquelle \bar{x}_g aide à prédire y_i quand on a déjà neutralisé x_i , c'est-à-dire la mesure dans laquelle un effet contextuel linéaire aide à prédire y_i sur les $i \in U$, en utilisant une régression par les moindres carrés pondérés avec N_g comme pondération.

Les preuves de théorèmes 1 et 2 dépendent fortement de l'hypothèse d'échantillonnage en grappes. On ne s'attendrait pas à ce que les résultats soient applicables en cas de sous-échantillonnage dans les ménages.

Les théorèmes 1 et 2 s'appliquent uniquement quand $a_g = 1$ dans la régression par les moindres carrés pondérés pour \hat{T}_h . D'autres choix de a_g sont souvent utilisés; par exemple, il serait raisonnable de supposer que $v_{g1} = N_g^{-1}$ dans le modèle (5), auquel cas il serait logique d'utiliser $a_g = N_g^{-1}$. Le théorème 3 montre que \hat{T}_h est équivalent à un estimateur GREG au niveau de la personne ajusté sous le modèle contextuel linéaire pour d'autres choix de a_g .

Théorème 3. L'estimateur GREG contextuel linéaire

Pour les plans d'échantillonnage où toutes les personnes sont sélectionnées dans les ménages échantillonnés et ces populations, en sélectionnant d'abord des ménages par échantillonnage aléatoire simple sans remise, puis tous les membres des ménages échantillonnés. Nous avons simulés 10 000 ménages. Dans chaque cas, 5 000 échantillons ont été sélectionnés. Les variables auxiliaires x_i correspondaient à des variables indicatrices du sexe selon le groupe d'âge (12 catégories). (Ce choix de x_i signifie que l'estimation GREG équivaut à une poststratification.) Nous avons calculé l'estimateur GREG au niveau de la personne avec $c_i = 1(\hat{T}_p)$, l'estimateur GREG au niveau du ménage avec $a_g = N_g^{-1}(\hat{T}_h)$, et l'estimateur GREG au niveau du ménage avec $a_g = 1(\hat{T}_h)$. Nous avons également inclus l'estimateur de Hajek

Le théorème 3 signifie que \hat{T}_h est l'estimateur GREG sous un modèle plus général que \hat{T}_p . Nieuwenbroek (1993) a montré que \hat{T}_h est égal à un estimateur GREG au niveau de la personne dérivé par régression de y_i sur \bar{x}_g . Le théorème 3 énonce qu'il est également égal à l'estimateur GREG au niveau de la personne provenant de la régression de y_i sur x_i ainsi que sur \bar{x}_g , donc qu'il intègre automatiquement tous effets contextuels du ménage. Par conséquent, \hat{T}_h devrait en principe avoir une variance plus faible que \hat{T}_p pour les grands échantillons. (Dans le cas $a_g = 1$, le théorème 1 énonçait que cela est toujours le cas.) Pour les petits échantillons, par contre, un modèle plus général est parfois contreproductif. Silva et Skinner (1997) ont montré, pour l'échantillonnage à un seul degré, que

$$\text{var}_p[\hat{T}_p] - \text{var}_p[\hat{T}_h] = \frac{M^2}{m} \left(1 - \frac{m}{M} \right) (M - 1)^{-1} B_C^T \left(\sum_{i \in U} x_{g1} x_{g1}^T \right) B_C$$

où \hat{T}_h est calculé en utilisant $a_g = 1$ pour tout g .

pour \hat{T}_p et

$$g_i = 1 + (T^X - \hat{T}^{X*})^T \left(\sum_{g \in \mathcal{G}} a_g^T \pi_{g1}^{-1} x_{g1} x_{g1}^T \pi_{g1}^{-1} \right)^{-1} a_g^T \pi_{g1}^{-1} x_{g1}$$

pour \hat{T}_H , où la personne i appartient au ménage g . (L'indice supérieur « - » indique l'inverse généralisée d'une matrice).

2.3 Résultats théoriques

À la présente section, nous montrons que \hat{T}_H possède la variance en grand échantillon la plus faible possible dans une classe d'estimateurs qui comprend aussi \hat{T}_p pour le plan de sondage où les ménages sont sélectionnés par échantillonnage aléatoire simple sans remise. Puis, nous expliquons ce résultat en montrant que \hat{T}_H est équivalent à un estimateur par la régression calculé en utilisant les données au niveau de la personne, où le modèle contient des effets contextuels.

Pour les grands échantillons, \hat{T}_p et \hat{T}_H peuvent être approximés par

$$\hat{T}_p = \hat{T}_p^* + B_p^T(T^X - \hat{T}^{X*});$$

et

$$\hat{T}_H = \hat{T}_H^* + B_H^T(T^X - \hat{T}^{X*})$$

respectivement, où B_p et B_H sont les solutions de

$$(8) \quad \begin{cases} \sum_{i \in U} c_i(Y_i - B_H^T x_i) x_i = 0 \\ \sum_{g \in \mathcal{G}} a_g^T(Y_{g1} - B_p^T x_{g1}) x_{g1} = 0 \end{cases}$$

(Samdal et coll., 1992, Résultat 6.6.1, page 235). Le théorème 1 énonce l'estimateur à variance minimale dans une classe incluant \hat{T}_p et \hat{T}_H .

Théorème 1. Estimateur optimal pour

l'échantillonnage en grappes simple

Supposons que m ménages soient sélectionnés par échantillonnage aléatoire simple sans remise à partir d'une population de M ménages, et que tous les membres des ménages échantillonnés soient sélectionnés. Considérons l'estimateur de T donné par

$$\hat{T} = \hat{T}_p^* + B^T(T^X - \hat{T}^{X*})$$

où B est un vecteur de dimension p constant. Nous supposons qu'il existe un vecteur λ tel que $\lambda^T x_i = 1$ pour tout $i \in U$. La variance de cet estimateur est minimisée par les valeurs B^* qui sont les solutions de

$$\sum_{g \in \mathcal{G}} (Y_{g1} - B^T x_{g1}) x_{g1} = 0.$$

$$(9) \quad \begin{cases} E_M[Y_i] = \gamma_1^T \bar{x}_g + \gamma_2^T x_i \quad (i \in U_g) \\ \text{var}_M[Y_i] = \sigma^2 \\ Y_i, Y_j \text{ indépendantes pour } i \neq j. \end{cases}$$

Le théorème 1 a l'implication peut-être étonnante que \hat{T}_H (avec $a_g = 1$ pour tout g) a une variance plus faible que \hat{T}_p pour les grands échantillons, et cela en dépit du fait que \hat{T}_H écarte une partie de l'information contenue dans l'échantillon, parce qu'il utilise les sommes de x_i et y_i sur l'ensemble des ménages. Le théorème donne à penser que \hat{T}_H est l'estimateur GREG approprié pour le plan d'échantillonnage en grappes supposé ici et que l'information écartée par sommation au niveau du ménage n'est pas pertinente quand on utilise ce plan. Pour expliquer pourquoi \hat{T}_H peut donner de meilleurs résultats que \hat{T}_p , nous utiliserons un « modèle contextuel linéaire » qui est un modèle plus général de $E_M[Y_i]$ que (4). Ce modèle est :

Le théorème 2 montre que l'amélioration de la variance due à l'utilisation de \hat{T}_H avec $a_g = 1$ plutôt que \hat{T}_p peut être expliquée par le modèle contextuel linéaire.

Théorème 2. Explication de la différence entre les variances asymptotiques

Supposons que les ménages soient sélectionnés par échantillonnage aléatoire simple sans remise et que tous les membres des ménages échantillonnés soient sélectionnés. Soit $y_i = y_1 - B_p^T x_i$, et soit B_c le résultat de la régression de y_i en fonction de \bar{x}_g sur les $i \in U$ par la méthode des moindres carrés pondérés en utilisant N_g comme pondération. Alors

grappes est robuste à l'erreur de spécification des corrélations intra-grappe. Une interprétation de ce résultat serait que les corrélations à l'intérieur du ménage ne sont pas pertinentes en ce qui concerne l'estimation des totaux de population, parce que tous les membres des ménages échantillonnés sont sélectionnés. Donc, les corrélations à l'intérieur des ménages ne facilitent pas l'estimation pour les personnes non échantillonnées, puisque les personnes échantillonnées et non échantillonnées appartiennent à des ménages distincts.

Un certain nombre de méthodes ont été proposées pour l'estimation de type GREG avec pondérations égales dans les ménages. Nieuwenbroek (1993) a proposé un estimateur dont la motivation était l'agrégation du modèle (4) au niveau du ménage :

$$(5) \quad \left\{ \begin{array}{l} E_M[Y_{gi}] = \mathbf{B}^T \mathbf{x}_{gi} \\ \text{var}_M[Y_{gi}] = v_{gi} \sigma^2 \\ y_{gi}, y_{li} \text{ indépendantes pour } g \neq l. \end{array} \right.$$

où $v_{gi} = \sum_{i \in U_g} v_i$. L'estimateur GREG calculé en utilisant les données d'échantillon y_{gi} pour $g \in s_1$ basé sur ce modèle est \hat{T}^H :

$$(6) \quad \hat{T}^H = \hat{T}^{\pi} + \hat{\mathbf{B}}_T^H (T^{\pi} - \hat{T}^{X^{\pi}}) \quad \text{où } \hat{\mathbf{B}}^H \text{ est une solution de} \\ (7) \quad \sum_{g \in s_1} \pi_{gi}^{-1} a_{gi}^g (v_{gi}^{-1} - \hat{\mathbf{B}}_T^H \mathbf{x}_{gi}^T) \mathbf{x}_{gi} = \mathbf{0}.$$

Le coefficient de régression $\hat{\mathbf{B}}^H$ est une régression par les moindres carrés pondés au niveau du ménage des valeurs d'échantillon de y_{gi} sur \mathbf{x}_{gi}^T avec les pondérations $\pi_{gi}^{-1} a_{gi}^g$. Les valeurs de a_{gi}^g pourraient être fixées à v_{gi}^{-1} . Si $v_i = 1$ alors $v_{gi} = N_{gi}^{\pi}$ de sorte que $a_{gi}^g = N_{gi}^{\pi}$. Sinon, $a_{gi}^g = 1$ pourrait également être utilisé.

Plusieurs autres méthodes à pondération intégrée équivalentes ont été utilisées. Lemaître et Dufour (1987) ont construit un estimateur par la régression généralisée au niveau de la personne en utilisant \mathbf{x}_i^g au lieu de \mathbf{x}_i comme variables auxiliaires. Nieuwenbroek (1993) a fait remarquer que cela est équivalent à (6) si $c_i = a_i^g N_{gi}^{\pi}$ pour $i \in U_g$. Alexander (1987) a élaboré des méthodes de pondération étroitement reliées en utilisant un critère de distance minimale.

Les estimateurs GREG aux niveaux de la personne et du ménage peuvent tous deux s'écrire sous la forme pondérée $\sum_{i \in s} w_i y_i$. Les pondérations pour les deux estimateurs peuvent s'écrire $w_i = \pi_i^{-1} g_i$ où

$$g_i = 1 + (T^X - \hat{T}^{X^{\pi}})^T \left(\sum_{i \in s} c_i \pi_i^{-1} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} c_i \mathbf{x}_i$$

2.2 Estimateurs GREG aux niveaux de la personne et du ménage

sous le plan pour \mathbf{B} . Coefficients de régression d'échantillon $\hat{\mathbf{B}}$ sont convergents moindres carrés pondés de y_i sur \mathbf{z}_i pour $i \in U$. Les coefficients \mathbf{B} sont calculés d'après une régression par les

$$\sum_{i \in U} c_i (y_i - \mathbf{B}^T \mathbf{z}_i) \mathbf{z}_i = \mathbf{0}$$

et \mathbf{B} est une solution de

$$\hat{T}^{\pi} = \hat{T}^{\pi} + \mathbf{B}^T (T^{\pi} - \hat{T}^{Z^{\pi}})$$

où

L'estimateur GREG au niveau de la personne, \hat{T}^p , est l'estimateur GREG sous le modèle suivant :

$$(4) \quad \left\{ \begin{array}{l} E_M[Y_i] = \mathbf{B}^T \mathbf{x}_i \\ \text{var}_M[Y_i] = v_i \sigma^2 \\ y_i, y_j \text{ indépendantes pour } i \neq j. \end{array} \right.$$

Donc, l'estimateur GREG au niveau de la personne, \hat{T}^p , est donné par substitution de \mathbf{x}_i à \mathbf{z}_i dans (2). Le modèle (4) ignore toute corrélation entre y_i et y_j pour les personnes i et j dans le même ménage. Ces corrélations étaient égales par Clark et Steel (2002), quoiqu'ils aient observé des valeurs plus élevées pour les variables associées à l'auto-identification en tant qu'Autochtone. Des corrélations de valeur 1 pourraient survenir pour des variables environnementales. Tam (1995) montre que l'estimateur assisté par modèle optimal pour l'échantillonnage en

article, nous nous concentrons sur la pondération intégrée et n'envisageons pas les approches d'étalement.

Luey (1986), Alexander (1987), Heidal (1992), ainsi que Lemaire et Dufour (1987) ont discuté d'un certain nombre de méthodes qui produisent des pondérations intégrées pour les estimations au niveau de la personne et au niveau du

ménage. Toutefois, aucun de ces auteurs n'a évalué l'effet sur la variance d'échantillonnage du calcul de l'estimateur par la régression généralisée au niveau du ménage plutôt qu'au niveau de la personne. Cette question est importante en pratique, car il convient de trouver le juste équilibre entre l'avantage cosmétique des pondérations intégrées et tout effet sur l'efficacité d'échantillonnage.

Le présent article donne une comparaison de la variance sous le plan, qui est la variance calculée par échantillonnage probabiliste répété à partir d'une population fixe, des estimateurs par la régression généralisée au niveau de la

personne et au niveau du ménage. À la section 2, nous prouvons que la variance en grand échantillon de l'estimateur au niveau de la personne, en montrant que le premier est optimal dans une grande classe d'estimateur GRÉG. Nous montrons qu'il en est ainsi parce que l'estimateur au niveau du ménage est inférieure ou égale à celle de simulation pour comparer les deux estimateurs en nous appuyant sur une gamme de variables. À la section 4, nous discutons des résultats. Les preuves des trois théorèmes sont présentées en annexe.

2. Comparaison théorique des estimateurs GRÉG aux niveaux de la personne et du ménage

2.1 L'estimateur par la régression généralisée

À la présente sous-section, nous décrivons l'estimateur par la régression généralisée pour le cas général de d unités. Soit U une population finie d'unités et $s \subseteq U$, l'échantillon. Les probabilités de sélection sont $\pi_i = \Pr\{i \in s\}$ pour les unités $i \in U$. Soit y_i la variable d'intérêt qui est observée pour les unités $i \in s$. Soit z_i le vecteur de variables auxiliaires pour l'unité i , qui sont observées pour chaque unité de la population. Les totaux de population de ces variables sont T_y et T_z , respectivement. L'estimateur par la régression généralisée de T_y est basé sur un modèle reliant la variable d'intérêt aux variables auxiliaires :

$$(1) \quad \begin{cases} E_M[y_i] = \beta_T' z_i \\ \text{var}_M[y_i] = v_i \sigma^2 \\ y_i, y_j \text{ indépendantes pour } i \neq j \end{cases}$$

où v_i représente les paramètres de variance connus. L'indice « M » désigne les espérances sous le plan, qui sont les espérances calculées par échantillonnage probabiliste répété à partir d'une population fixe. Dans le cas des enquêtes-entreprises, qui recueillent des données sur des variables continues telles que les revenus et les dépenses de l'entreprise, v_i est souvent modélisée sous forme d'une fonction de la taille de l'entreprise. Dans le cas des enquêtes-ménages, la variable d'intérêt est fréquemment dichotomique, auquel cas v_i est habituellement fixée à 1, ce qui correspond à un modèle homoscédastique.

Habituellement, z_i a la propriété qu'il existe un vecteur λ tel que $\lambda_T' z_i = 1$ pour tout $i \in U$. Par exemple, cela est vrai si le modèle de régression (1) contient un paramètre d'ordonnée à l'origine.

Définition 1. Estimateur par la régression généralisée L'estimateur par la régression généralisée pour le modèle (1) est défini comme étant

$$(2) \quad \hat{T}_y = \hat{T}_y^* + \hat{\beta}_T'(T_z - \hat{T}_z^*)$$

ou

$$\hat{T}_y^* = \sum_{i \in s} \pi_i^{-1} y_i$$

$$\hat{T}_z^* = \sum_{i \in s} \pi_i^{-1} z_i$$

et $\hat{\beta}$ est une solution de

$$\sum_{i \in s} c_i \pi_i^{-1} (y_i - \beta_T' z_i) z_i = 0$$

où c_i représente les poids de régression. (Souvent, c_i est fixé à $c_i = v_i^{-1}$.)

Les coefficients $\hat{\beta}$ sont calculés d'après une régression par les moindres carrés pondérés de y_i sur z_i pour $i \in s$. L'estimateur GRÉG possède une variance sous le plan faible si le modèle est approximativement vrai, mais est convergent sous le plan indépendamment de la vérité du modèle (par exemple Samdal et coll. 1992, chapitre 6). Pour de grands échantillons, la variance sous le plan de \hat{T}_y est approximativement égale à

$$(3) \quad \text{var}_p[\hat{T}_y] \approx \text{var}_p[\hat{T}_y^*]$$

Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes-ménages

David G. Steel et Robert G. Clark¹

Résumé

Une classe courante de plans de sondage comprend la sélection de toutes les personnes dans les ménages échantillonnés. Des estimateurs par la régression généralisée peuvent être calculés au niveau de la personne ou du ménage. L'utilisation de cas de l'échantillonnage aléatoire simple de ménages et la sélection de toutes les personnes présentes dans chaque ménage membres du ménage. Dans le présent article, nous comparons théoriquement et empiriquement les deux approches dans le cas de l'échantillonnage. Nous constatons que l'approche au niveau du ménage est théoriquement plus efficace dans le cas de grands échantillons et que toute inefficacité empirique dans les petits échantillons est limitée.

Mots clés : Effets contextuels; estimateur par la régression généralisée; corrélation intraclass; variance d'échantillonnage; assisté par modèle; enquêtes-ménages.

1. Introduction

De nombreuses enquêtes-ménages comprennent la sélection d'un échantillon de ménages, suivie de la sélection de toutes les personnes faisant partie du champ d'observation de l'enquête dans les ménages échantillonnés. Des données sur une ou plusieurs variables d'intérêt sont recueillies pour les personnes incluses dans l'échantillon. Il existe parfois des variables auxiliaires pour lesquelles les totaux de population et les valeurs d'échantillon sont connues; par exemple, il pourrait s'agir de chiffres de population selon les caractéristiques géographiques et démographiques. L'estimateur par la régression généralisée (GREG) est souvent utilisé pour combiner l'information auxiliaire et les données d'échantillon en vue d'estimer efficacement les totaux de population de la variable d'intérêt. L'estimateur GREG s'appuie sur un modèle de régression reliant la variable d'intérêt aux variables auxiliaires. L'approche ordinaire consiste à ajuster ce modèle en utilisant les données recueillies pour chaque personne faisant partie de l'échantillon (par exemple Lemaître et Dufour 1987, premier paragraphe). Cet estimateur GREG au niveau de la personne est égal à une somme pondérée des valeurs d'échantillon de la variable d'intérêt, où la pondération est en général différente pour chaque personne. Il est parfois comme mode d'utiliser des pondérations égales pour tous les membres d'un ménage dans le cas d'enquêtes qui recueillent des renseignements sur des variables d'intérêt au niveau du ménage ainsi qu'au niveau de la personne. Les mêmes pondérations peuvent alors être utilisées pour les deux types de variables afin d'être certain que les relations entre les variables du ménage et les variables personnelles soient reflétées dans les estimations

du total. Si une variable au niveau du ménage est égale à la somme des variables au niveau de la personne (par exemple, si le revenu du ménage est égal à la somme des revenus personnels), alors le total estimé de la variable au niveau du ménage sera égal au total estimé de la variable au niveau de la personne. Cela n'est généralement pas le cas si des méthodes de pondération distinctes sont utilisées pour les variables au niveau de la personne et au niveau du ménage. De même, s'il existe une inégalité entre la variable au niveau du ménage et la somme des variables au niveau de la personne, celle-ci sera reflétée dans les estimations des deux variables. Par exemple, le nombre estimé de ménages utilisant une garderie ne devrait pas être supérieur au nombre estimé d'enfants allant en garderie. L'estimateur GREG au niveau du ménage produit des pondérations égales au sein des ménages en ajustant le modèle de régression d'après les totaux au niveau du ménage de la variable d'intérêt et des variables auxiliaires (par exemple, Nieuwenbroek 1993). Les pondérations ayant cette propriété sont appelées pondérations intégrées. Une autre approche consisterait à utiliser des méthodes d'estimation différentes pour les variables au niveau du ménage et celles au niveau de la personne, puis à faire une correction pour forcer les estimations qui devraient être égales à concorder. Cette approche pour obtenir la cohérence entre les estimations calculées d'après des enquêtes-entreprenues annuelles et infra-annuelles (par exemple, Cholette 1984). L'application de l'approche d'étalonnage aux variables au niveau du ménage et au niveau de la personne des enquêtes-ménages nécessiterait l'identification explicite des variables aux niveaux de la personne et du ménage pour lesquelles les totaux de population devraient être égaux. Dans le présent

Tableau 2 Prix relatifs de deux modèles de voitures selon le Kelly Blue Book et taux de dépréciation moyen après repondération

Année	Incluant mises au rancard		Taux moyens de dépréciation		Poids <i>Ex post</i>
	Buick	Chrysler	Buick	Chrysler	
1	0,8622	0,1367	0,1743	2,5714	Chrysler
2	0,7361	0,1377	0,1753	2,5714	
3	0,6195	0,5420	0,1378	1,2857	2,5714
4	0,5092	0,4261	0,1379	1,2857	2,5714
5	0,4042	0,3234	0,1387	1,2857	2,5714
6	0,3058	0,2341	0,1404	1,2857	2,5714
7	0,2181	0,1597	0,1432	1,2857	1,2857
8	0,1441	0,1009	0,1475	1,2857	0,2338
9	0,0867	0,0580	0,1537	0,2571	0,2338
10	0,0448	0,0287	0,1654	0,2571	0,2338
11	0,0223	0,0137	0,1716	0,2571	0,2338
12	0,0094	0,0055	0,1824	0,2571	0,2338
13	0,0035	0,0019	0,1932	0,2571	0,2338
14	0,0012	0,0007	0,1999	0,2571	0,2338
15	0,0004	0,0002	0,2050	0,2571	0,2338
16	0,0001	0,0000	0,2194	0,2571	0,2338
17	0,0000	0,0000	0,2432	0,2571	0,2338
18	0,0000	0,0000	0,2542	0,2571	0,2338
Moyenne pondérée				0,1479	0,1836

Remerciements

Les auteurs tiennent à remercier grandement l'arbitre anonyme de *Techniques d'enquête* qui, par ses judicieux commentaires, a contribué à améliorer la qualité de l'article.

Bibliographie

Bickel, P.J., et Doksum, K.A. (1977). *Mathematical Statistics*, Holden-Day, Oakland, CA.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Davidsen, R., et MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, N.Y.

Ross, S.M. (2002). *Introduction to Probability Models*, 8^{ème} Edition, Academic Press.

Lancaster, T. (1985). Generalized residuals and heterogeneous duration model: With applications to the weibull model. *Journal of Econometrics*, 28, 155-69.

Huilen, C.R., et Wykoff, F.C. (1981). The measurement of economic depreciation. Dans *Depreciation, Inflation, and the Taxation of Income from Capital*, (Ed. C.R. Huilen). The Urban Institute Press, Washington, D.C, 81-125.

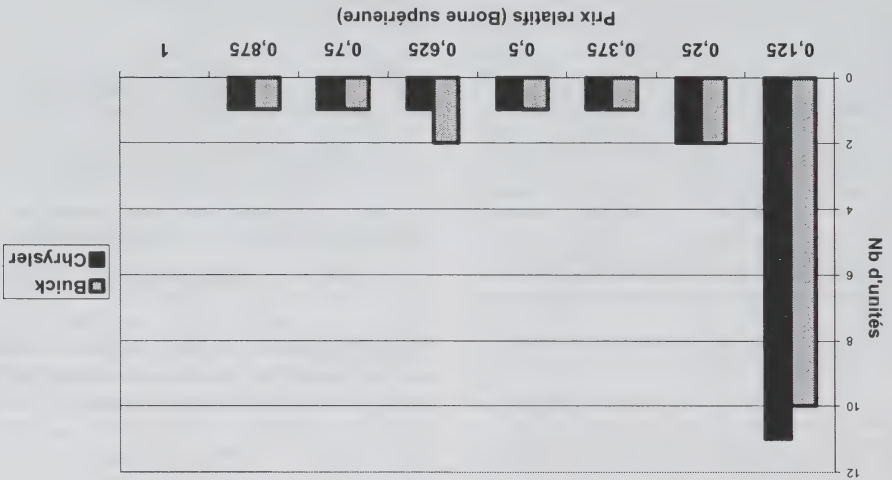
Greene, W.H. (1993). *Econometric Analysis*. Deuxième édition, Prentice Hall, Englewood Cliffs, N.J.

Gellatly, G., Tanguay, M. et Yan, B. (2002). An alternative methodology for estimating economic depreciation: New results using a survival model. Dans *Productivity Growth in Canada 2002*, Statistique Canada. #15-204-XPE.

Tableau 1 Prix relatifs de deux modèles de voitures selon le Kelly Blue Book et Taux de dépréciation moyen avant remédiation

Année	Pr (t > s)*			Taux moyens de dépréciation		
	Prix relatifs			Incluant mises au rancard		
	Excluant mises au rancard			Incluant mises au rancard		
	Buick	Chrysler	Buick	Chrysler	Buick	Chrysler
1	0,9988	0,8633	0,8257	0,8622	0,8246	0,1743
2	0,9901	0,7435	0,6801	0,7361	0,6734	0,1753
3	0,9666	0,6410	0,5608	0,6195	0,5420	0,1754
4	0,9220	0,5523	0,4621	0,5092	0,4261	0,1755
5	0,8526	0,4740	0,3794	0,4042	0,3234	0,1762
6	0,7582	0,4034	0,3087	0,3058	0,2341	0,1779
7	0,6433	0,3391	0,2482	0,2181	0,1597	0,1805
8	0,5164	0,2790	0,1953	0,1441	0,1009	0,1846
9	0,3892	0,2227	0,1491	0,0867	0,0580	0,1906
10	0,2731	0,1639	0,1050	0,0448	0,0287	0,2018
11	0,1770	0,1261	0,0772	0,0223	0,0137	0,2077
12	0,1051	0,0892	0,0523	0,0094	0,0055	0,2180
13	0,0567	0,0614	0,0344	0,0035	0,0019	0,2284
14	0,0276	0,0441	0,0236	0,0012	0,0007	0,2347
15	0,0120	0,0320	0,0164	0,0004	0,0002	0,2396
16	0,0046	0,0190	0,0093	0,0001	0,0000	0,2534
17	0,0016	0,0088	0,0041	0,0000	0,0000	0,2761
18	0,0005	0,0051	0,0023	0,0000	0,0000	0,2867
Moyenne	0,1727			0,1727		0,2087

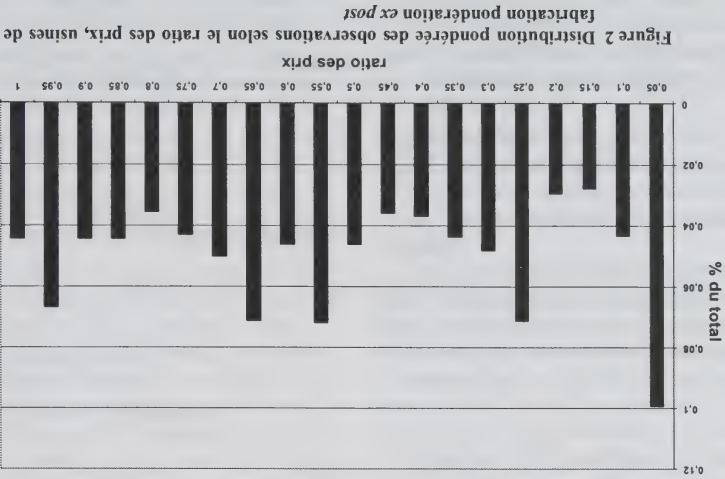
* Probabilité de Survie selon les estimations de la Division des études et de l'analyse micro-économique de Statistique Canada.



l'on utilise une simple moyenne des âges dans le calcul de \bar{t} revient à accorder de façon implicite le même poids à chacun des âges. Mais il est bien évident que ce ne serait pas la distribution que l'on obtiendrait si on tirait un échantillon aléatoire des voitures en services. La figure ci-dessous présente la distribution des cellules de prix entre les ratios de 0 et 1.

La technique de pondération consiste simplement à imposer un poids égal à chacune des fourchettes de prix relatifs. Dans cet exemple, les $n = 18$ âges sont répartis en $H = 7$ classes, ce qui répartit les âges en $18/7$ à chacune d'entre elles (en réalité, la structure des cellules a été configurée pour 8 classes mais la dernière est toujours vide). Comme mentionné à la section 3, les poids individuels w_i de chaque âge i sont construits selon (3), c'est-à-dire en divisant $18/7$ par le nombre d'observations qui se trouvent dans chaque classe, sauf pour les cellules vides dont le poids demeure nul. Le tableau 2 présente les résultats et l'impact de la repondération sur les statistiques dérivées.

Cet exemple illustre bien les problèmes de biais d'aggrégats économiques, sans tenir compte de la distribution réelle des unités au niveau micro. Ainsi, il est assez évident que les unités de 17 et 18 ans ne sauraient avoir le même poids de régression que celles de 1 an puisque le risque de perte à 1 an concerne pratiquement toutes les voitures qui seront exposées au risque de perte de valeur à des âges avancés. Il en résulte que l'estimation non pondérée, dans cet exemple, introduit une surestimation du taux de dépréciation de l'ordre de 15 %.



4. Application

Si on reprend l'histogramme exposé plus haut et divisons l'échantillon en $H = 5$ intervalles d'une largeur de 0,2 avec une valeur de $\pi = 1/5 = 0,2$, on obtient alors l'histogramme suivant qui a été pondéré *ex post*.

Nous allons illustrer la démarche à partir d'un exemple tiré du Kelly Blue Book, une source d'information largement utilisée pour l'estimation de la dépréciation des voitures. Le tableau 1 présente les prix de deux modèles de voitures pour différents âges entre 1 et 18 ans. Pour chaque voiture, on dispose donc d'un échantillon de $n = 18$ unités. Les prix sont exprimés en valeur relative par rapport à un modèle neutre. Il est en outre nécessaire d'ajuster les ratios pour tenir compte de la probabilité de survie à chacun de ces âges. Pour chaque voiture, le ratio final utilisé r_i de l'année i est donc construit à partir du produit du ratio des prix par la probabilité de survie.

On s'intéresse au taux de dépréciation moyen \bar{t} pour chaque voiture. Ce dernier pourrait être estimé à partir d'une régression des prix (ou d'une fonction de ces derniers) par rapport à l'âge (ou d'une fonction de l'âge). Toutefois, si on présuppose que le taux est constant et de forme géométrique, on a la relation $r_i = 1 - \bar{t}^i$, où r_i est le prix relatif selon l'âge i . Dans ce cas, un taux \bar{t}_i peut être estimé à chaque âge i par $\bar{t}_i = 1 - r_i^{1/i}$. Une estimation du taux de dépréciation pour tous les âges, c'est-à-dire $\bar{t} = \sum_{i=1}^{18} \bar{t}_i / 18$.

Dans l'exemple ci-dessus, on constate que les taux de dépréciation \bar{t}_i varient selon la fourchette d'âge et qu'ils ont tendance à augmenter avec l'âge. Par ailleurs, le fait que

Il est facile de vérifier ces résultats de façon numérique, à l'aide de données simulées et nous ne nous y attarderons pas. Nous allons plutôt examiner comment ce résultat peut être réintroduit dans la base de données pour lui restaurer, au moins en partie, des propriétés semblables à celle d'un échantillon aléatoire. Pour ce faire, il nous suffit d'imposer *ex post*, à la distribution empirique des prix, une structure de poids w_i qui soit telle que la distribution empirique des données, dans l'espace des prix, soit uniforme.

par

$$F^n(y) = \frac{n}{\sum_{i=1}^n I_i(y)}$$

(1)

où $I_i(y) = 1$ si la valeur mesurée r_i de l'actif i est inférieure

au total d'observations. Notons que si les n unités de l'échantillon sont indépendantes et identiquement distribuées (i.i.d.), lorsque $n \rightarrow \infty$, $F^n(y)$ converge en probabilité vers $F(y)$, c'est-à-dire $F^n(y) \xrightarrow{P} F(y)$ (Bickel et Doksum 1977).

Pour obtenir le poids w_i pour chaque actif i , on distribue

simplement l'échantillon en un nombre H donné d'intervalles (ou classes) de largeur fixe sur l'échelle des ratios de prix, et on attribue la même probabilité $\pi = 1/H$ à chacun de ces intervalles. Puisque les ratios de prix sont contenus entre 0 et 1, on a alors l'intervalle $h = 1$ donné par

[0, $(h-1)H^{-1}$, hH^{-1}]. Un poids w_h est ensuite calculé dans chaque intervalle h par le ratio π/π_h où π_h est la probabilité empirique spécifique à l'intervalle h , c'est-à-dire

$$\pi_h = \frac{1}{n} \sum_{i=1}^n \delta_i(h) = \frac{n}{Hn_h}$$

où $\delta_i(h) = 1$ si $r_i \in h$, 0 sinon. On pose alors

$$w_h = \frac{\pi}{\pi_h} = \frac{n}{Hn_h}$$

pour $r_i \in h$. En utilisant ces poids, la distribution pondérée empirique des ratios de prix r est donnée par

$$F^{n,w}(y) = \frac{\sum_{i=1}^n w_i I_i(y)}{\sum_{i=1}^n w_i}$$

(4)

En notant que $\sum_{i=1}^n w_i = \sum_{h=1}^H n_i/Hn_h = n$, on obtient finalement

$$F^{n,w}(y) = \frac{n}{\sum_{i=1}^n w_i I_i(y)} \quad (5)$$

Puisque $n_h = \sum_{i=1}^n \delta_i(h)$, on a

$$F^{n,w}(y) = \frac{\sum_{i=1}^n w_i I_i(y)}{n} = \frac{\frac{1}{H} \sum_{h=1}^H \frac{H}{n_h} \sum_{i=1}^n \delta_i(h) I_i(y)}{\frac{1}{H} \sum_{h=1}^H \frac{H}{n_h} \sum_{i=1}^n \delta_i(h) I_i(y)} = \frac{\frac{1}{H} \sum_{h=1}^H \frac{H}{n_h} \sum_{i=1}^n \delta_i(h) I_i(y)}{\frac{1}{H} \sum_{h=1}^H \frac{H}{n_h} \sum_{i=1}^n \delta_i(h) I_i(y)}$$

(6)

Lorsque $n \rightarrow \infty$, on a $(1/n) \sum_{i=1}^n \delta_i(h) I_i(y) \xrightarrow{P} P(r \in h, r \leq y)$ et $(1/n) \sum_{i=1}^n \delta_i(h) \xrightarrow{P} P(r \in h)$. Donc,

$$F^n(y|h) \xrightarrow{P} \frac{P(r \in h, r \leq y)}{P(r \in h)}$$

où $F(y|h)$ est la distribution des ratios de prix r à l'intérieur de l'intervalle h .

Pour n suffisamment grand, on devrait être en mesure de déterminer H de manière à construire les intervalles h pour

que $F^n(y|h)$ soit approximativement uniformément distribuée, $h = 1, \dots, H$. En d'autres mots, lorsque $n \rightarrow \infty$, pour H suffisamment grand, $F(y|h)$ devrait suivre une distribution uniforme sur l'intervalle h . Notons qu'un tel argument a été utilisé par Daley et Hodges (1959) dans un contexte de stratification optimale. Dans ce cas, la distribution $F(y|h)$ est donnée par

$$F(y|h) = \begin{cases} 0 & \text{pour } y \leq (h-1)H^{-1} \\ 1 & \text{pour } y > hH^{-1} \end{cases} \quad (8)$$

Puisque $F(y) = \sum_{h=1}^H F(y|h)/H$, on a $F(y) = y$, ce qui correspond à la distribution uniforme. On conclue donc que pour n suffisamment grand, l'utilisation de la pondération (3) devrait rendre la distribution empirique pondérée $F^{n,w}(y)$ donnée par (5) approximativement uniformément distribuée.

Des simulations Monte-Carlo ont démontré que les estimations résultant d'un échantillon non aléatoire pouvaient être améliorées en utilisant cette approche. Ses principaux avantages résident dans :

- sa simplicité;
- le fait qu'elle peut être introduite *ex ante*, c'est-à-dire avant l'introduction du modèle économique comme tel. Par conséquent, elle ne requiert pas d'hypothèses de travail fortes.

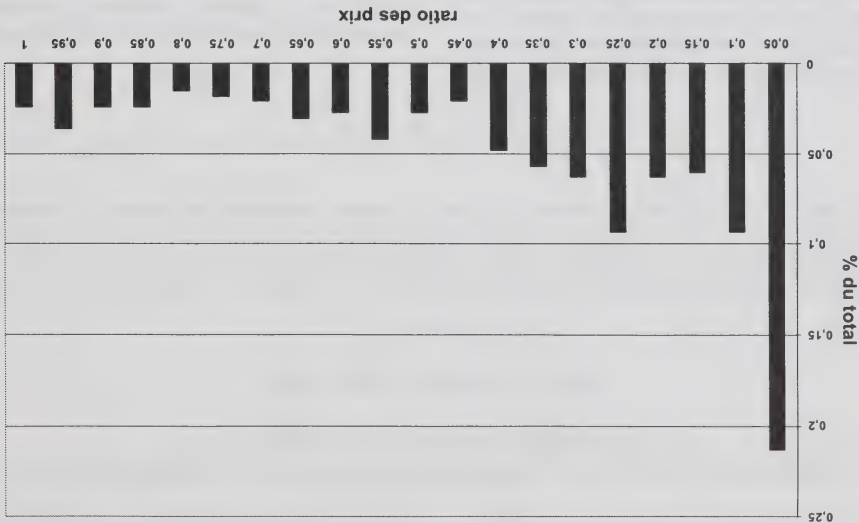


Figure 1 Distribution des observations selon le ratio des prix, usines de fabrication

3. Démarche

Note: point de départ est que les ratios de prix peuvent être considérés comme des réalisations empiriques d'une

fonction de survie de forme inconnue. Dans les modèles de durée de vie, la fonction de survie exprime la probabilité qu'une entité dont la vie est limitée survive au-delà d'un certain point sur l'axe du temps. Elle fournit, par conséquent, la même information que la fonction de répartition (ou *Cumulative Distribution Function*). Soit r , une variable aléatoire qui décrit la durée de vie d'une unité de valeur incorporée dans un actif quelconque. La valeur s'épuise au fur et à mesure que le temps passe, et ce, aussi longtemps que l'actif est en service. Le ratio des prix peut donc s'interpréter comme la fraction survivante qui diminue peu à peu. On note cette fraction $S(y)$ et on a

$$S(y) = 1 - F(y)$$

où $F(y) = P(r \leq y)$ est la fonction de répartition, c'est-à-dire la probabilité qu'une unité de valeur soit perdue avant le point y .

Les théorèmes des transformations fondamentales des lois de probabilité permettent de décrire la fonction inverse de $F(y)$ (Greene 1993 et Ross 2002). Soit $z = F(y)$ et supposons que la fonction inverse F^{-1} existe de sorte que $y = F^{-1}(z)$. Il y a donc une concordance directe entre l'espace de y , borné à 0 mais infini à droite, et celui de F qui

Dans le cas des données de prix, l'intuition est la

suite: entre l'investissement et la mise au rancard, toute la fourchette des prix relatifs doit forcément être couverte par un actif en production. À la période initiale, la valeur perdue plus rapidement, il y a donc une plus grande quantité d'observations dont les durées sont courtes. Mais cela est compensé par le fait que la référence correspondante sur l'échelle du temps est également plus courte. Par exemple, il faut moins de temps pour passer de 100 % de la valeur initiale à 90 %, que de 15 % à 5 % de la valeur initiale.

La loi qui génère une telle répartition est une distribution uniforme entre 0 et 1. Ce résultat est généralement au cœur des processus de génération de données comme les simulations Monte-Carlo puisque lors de la génération d'un échantillon aléatoire, on utilise souvent une distribution uniforme à laquelle on applique ensuite la fonction inverse (Davidson et MacKinnon 1993). Cette approche n'est toutefois par toujours pratique ou carrément impossible, en particulier si la fonction inverse F^{-1} n'a pas de forme explicite. Ce résultat a aussi été utilisé dans les approches de résidus généralisés, notamment pour la construction de tests de spécification (Lancaster 1985).

Il en résulte que n'importe quel échantillon aléatoire construit à partir de réalisations empiriques de données de proportion de survie doit converger en distribution vers une

Pondération *ex post* des données de prix pour l'estimation des taux de dépréciation

Marc Tanguay et Pierre Lavallée¹

Résumé

Pour modéliser la dépréciation économique, on utilise une base de données qui contient des informations sur les actifs dont des entreprises se départissent. On connaît les prix d'acquisition et de vente ainsi que les durées d'utilisation de ces actifs. Cependant, les actifs dont on observe les prix sont uniquement ceux qui ont fait l'objet d'une transaction. Bien que la dépréciation d'un actif soit présentée de façon continue au cours de sa vie, on ne connaît donc cette valeur que lorsqu'il y a eu une transaction. La présente note propose une pondération *ex post* afin d'atténuer, au moins en partie, cet effet dans la détermination des modèles économétriques.

Mots clés : Ratio de prix; données de survie; distribution uniforme; dépréciation des voitures.

1. Contexte

Différents modèles économétriques sont utilisés pour estimer la dépréciation économique. On utilise, à cette fin, une base de données qui contient des informations sur les prix dont des entreprises se départissent. On connaît les prix d'acquisition et de vente ainsi que les durées d'utilisation de ces actifs. On voudrait en inférer les résultats à la population totale des actifs utilisés par les entreprises. Sur l'utilisation de prix d'actifs usagés pour estimer la dépréciation économique, on peut notamment consulter Gellatly, Tanguay et Van (2002), ainsi que Hulten et Wykoff (1981).

On s'interroge cependant sur la représentativité de la base de données utilisée. En effet, les actifs dont on observe les prix sont uniquement ceux qui ont fait l'objet d'une transaction. On ignore dans quelle mesure les pertes de valeur observées sur eux sont représentatives de la perte de valeur pour tous les actifs en production, qu'ils aient ou non fait l'objet d'une transaction. Ceci peut constituer une source d'erreur dans l'établissement des modèles économétriques parce que ces derniers cherchent à mesurer la dépréciation des actifs au cours de leur vie, qu'il y ait eu transaction ou non.

Nous nous proposons d'atténuer cette source d'erreur, en introduisant une pondération *ex post* dans la détermination des modèles économétriques. La section 2 de la présente note décrira plus en détail la problématique. À la section 3, nous exposerons la démarche suivie pour la détermination de la pondération. Finalement, à la section 4, nous présenterons quelques résultats numériques.

2. Problématique

On cherche à décrire la relation entre les prix et l'âge des actifs. On dispose d'un échantillon de n actifs où on connaît, pour chaque actif i , le ratio de prix r_i et le temps t_i où ce

On peut donc se demander quelle forme aurait la distribution ci-dessus si elle avait été tirée d'un échantillon où le ratio des prix avait été mesuré, pour un même actif i , à différents temps t . Notre argument est qu'elle devrait converger vers une *distribution uniforme*. Nous allons donc chercher à obtenir une pondération qui nous aidera à recréer une distribution uniforme des ratios de prix. Cette pondération nous aidera à pallier le manque d'uniformité dans la distribution des observations, ce qui peut influencer les analyses statistiques comme, par exemple, la régression linéaire.

Étant donné que l'on veut utiliser les données pour inférer des statistiques sur la population des actifs en production, on souhaiterait que nos données aient des propriétés analogues à celles d'un échantillon aléatoire qui serait tiré sur cette population. Ceci n'est pas le cas, rappelons-le, parce qu'on ne dispose que des prix des actifs i qui ont fait l'objet d'une transaction au temps t_i , $i = 1, \dots, n$. En effet, bien que l'on voudrait disposer de ratios des prix pour différentes périodes de l'existence d'un actif i donné, ce ratio n'est disponible que lorsqu'il y a eu transaction, ce qui survient de façon non uniforme durant la durée de vie d'un actif.

Une fois que les prix sont exprimés en dollars réels, ce ratio est donné par $r_i = P_t / P_0$ où P_0 est la valeur initiale de l'investissement de l'actif i et P_t est son prix de vente au temps t . Ce ratio est strictement décroissant par rapport à l'axe du temps t . Au point de départ, on ignore le processus qui génère la perte de valeur et on n'a aucune spécification concernant la fonction qui décrit cette perte, sinon qu'elle est strictement décroissante. Il est cependant possible d'examiner la distribution des ratios des prix entre 0 et 1. Voici un exemple construit à partir des données sur les usines de fabrication (on doit noter que les 2/3 de l'échantillon ont été exclus car il correspondait à des mises au rancard (le prix est nul) et les procédures d'estimation prennent en compte, chacune à sa façon, cette composante).

1. Marc Tanguay et Pierre Lavallée, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Courriel : marc.tanguay@statcan.ca, pierre.lavallee@statcan.ca.

Preuve du théorème 3.2 : Notons que \hat{y}_{diff} peut s'écrire comme étant la somme d'une constante de population et d'un estimateur à facteur d'extension de la forme $\hat{y}_k - s_{\text{rk}}^T Y_U + s_{\text{rk}}^k Z_U^T B - z_k B$ pour $k \in U$. Comme dans le cas de la variable y_k , cette nouvelle variable a un support borné en vertu du lemme 1(b) et une variance d'ordre $O(1/n)$ en vertu du lemme 4. Donc, l'existence du TLC pour \hat{y}_π implique l'existence du TLC pour \hat{y}_{diff} . En outre, $\hat{y}_{\text{reg}}^{\text{reg}} = \hat{y}_{\text{diff}} + o_p(1/\sqrt{n})$ en vertu du lemme 3, de sorte que $\sqrt{n} \hat{y}_{\text{reg}}^{\text{reg}}$ et $\sqrt{n} \hat{y}_{\text{diff}}^{\text{diff}}$ suivent la même loi asymptotique. L'application du théorème de Slutsky et du lemme 5 complète la preuve.

Bibliographie

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.

Francisco, C.A., et Fuller, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2^{ème} Ed.). New York : John Wiley & Sons, Inc.

Hastie, T.J., et Tibshirani, R.J. (1990). *Generalized Additive Models*. Washington, D.C.: Chapman and Hall.

Isaki, C., et Fuller, W. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Larsen, D.P., Thornton, K.W., Urquhart, N.S. et Paulsen, S.G. (1993). Overview of survey design and lake selection. EMAP - Surface Waters 1991 Pilot Report. (Eds. D.P. Larsen et S.J. Christie). Rapport Technique EPA/620/R - 93/003, U.S. Environmental Protection Agency.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, 73, 166-179.

Opsomer, J.D., Breidt, F.J., Moisen, G.G. et Kauermann, G. (2007). Model-assisted estimation of forest resources with generalized additive models. *Journal of the American Statistical Association*. A paraître.

Opsomer, J.-D., et Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186-211.

Opsomer, J.D., et Ruppert, D. (1999). A root-n consistent estimator for semiparametric additive modelling. *Journal of Computational and Graphical Statistics*, 8, 715-732.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Speckman, P.E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society, Series B*, 50, 413-436.

Stoddard, J.L., Kahl, J.S., Deviney, F.A., DeWalle, D.R., Driscoll, C.T., Herlihy, A.T., Kellogg, J.H., Murdoch, P.S., Webb, J.R. et Webster, K.E. (2003). Response of surface water chemistry to the Clean Air Act Amendments of 1990. Rapport Technique EPA/620/R-03/001, U.S. Environmental Protection Agency.

U.S. National Acid Precitation Assessment Program (1991, Novembre). 1990 Integrated Assessment Report. Rapport Technique, Washington, DC.

$$(16) \quad \left(\bar{\mathbf{z}}_n - \bar{\mathbf{z}}_n^* \right) (\mathbf{B} - \mathbf{B}) = o_p \left(\frac{\sqrt{n}}{1} \right)$$

$$(17) \quad \frac{1}{n} \sum_{i=1}^n (m_k - m_k^*) \left(1 - \frac{\pi_k}{I_k} \right) = o_p \left(\frac{\sqrt{n}}{1} \right).$$

Le lemme 2 et les hypothèses A2, A5 et A6 montrent que $(\bar{\mathbf{z}}_n^* - \bar{\mathbf{z}}_n^*)(\mathbf{B} - \mathbf{B}) = o_p(1/nh)$. Afin de prouver (17), nous pouvons réécrire cette équation sous la forme

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (m_k - m_k^*) \left(1 - \frac{\pi_k}{I_k} \right) &= \frac{1}{n} \sum_{i=1}^n \left(\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{y}}_{[s]_1}^k \right) \left(1 - \frac{\pi_k}{I_k} \right) \\ &- \frac{1}{n} \sum_{i=1}^n \left(\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{z}}_{[s]_1}^k \right) \left(1 - \frac{\pi_k}{I_k} \right) (\mathbf{B} - \mathbf{B}). \end{aligned}$$

Breidt et Opsomer (2000) ont prouvé dans le lemme 5 que le premier terme du deuxième nombre est $o_p(1/\sqrt{n})$; ce même lemme et la limitabilité de \mathbf{B} fournissent le même taux pour le deuxième terme. Les hypothèses A5 et A6, le lemme 1(b) et le lemme 2 montrent que le troisième terme est $O_p(1/n\sqrt{h})$ et le taux souhaité est obtenu.

Lemme 4 Sous les hypothèses A6 et A8,

$$\begin{aligned} E^p(\hat{\mathbf{y}}_{\text{dir}}^p) &= \bar{\mathbf{y}}_n^p \\ \text{Var}^p(\hat{\mathbf{y}}_{\text{dir}}^p) &= \frac{1}{N^2} \sum_{k,l \in I} (\pi_k - \pi_k \pi_l) \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l} = o \left(\frac{n}{1} \right). \end{aligned}$$

Preuve du lemme 4 : Les propriétés de l'estimateur par la différence sont calculées facilement. Le taux de la variance utilisant le même raisonnement que pour le lemme 4 de Breidt et Opsomer (2000).

Lemme 5 Sous les hypothèses A1 à A8,

$$V(\hat{\mathbf{y}}^{\text{reg}}) = \text{Var}^p(\hat{\mathbf{y}}_{\text{dir}}^p) + o_p \left(\frac{n}{1} \right).$$

Preuve du lemme 5 : Le raisonnement de cette preuve suivra étroitement celui du théorème 3 de Breidt et Opsomer (2000). Nous écrivons

$$V(\hat{\mathbf{y}}^{\text{reg}}) - \text{Var}^p(\hat{\mathbf{y}}_{\text{dir}}^p) = (V(\hat{\mathbf{y}}^{\text{reg}}) - V(\hat{\mathbf{y}}_{\text{dir}}^p)) + (V(\hat{\mathbf{y}}_{\text{dir}}^p) - \text{Var}^p(\hat{\mathbf{y}}_{\text{dir}}^p)) \quad (18)$$

avec

$$\frac{1}{N^2} \sum_{k,l} \sum_{i=1}^n \pi_k \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l}.$$

Puisque

$$\frac{1}{n} \sum_{i=1}^n (\gamma_k - g_k)^4 < \infty.$$

en vertu des hypothèses A1 à A3 et d'après les lemmes 1(b) et 2, l'approche utilisée pour le terme $\hat{\mathbf{y}}_n^{\text{reg}}$ de Breidt et Opsomer (2000) peut servir à montrer que

$$E^p |V(\hat{\mathbf{y}}_{\text{dir}}^p) - \text{Var}^p(\hat{\mathbf{y}}_{\text{dir}}^p)| = o \left(\frac{n}{1} \right),$$

qui fournit la convergence souhaitée par l'inégalité de Markov.

Pour le premier terme de (18), notons que

$$\begin{aligned} g_k - g_k^* &= (\bar{\mathbf{y}}_{[s]_1}^k - \bar{\mathbf{y}}_{[s]_1}^k) - (\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{z}}_{[s]_1}^k) (\mathbf{B} - \mathbf{B}) \\ &+ (\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{z}}_{[s]_1}^k) (\mathbf{B} - \mathbf{B}) - (\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{z}}_{[s]_1}^k) \mathbf{B}, \end{aligned}$$

de sorte que

$$\begin{aligned} V(\hat{\mathbf{y}}^{\text{reg}}) - V(\hat{\mathbf{y}}_{\text{dir}}^p) &= \frac{1}{N^2} \sum_{k,l} \left\{ -2 \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l} \right. \\ &\quad \left. + \frac{\pi_k}{g_k - g_k^*} \frac{\pi_l}{g_l - g_l^*} \right\} \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l} I_k I_l \end{aligned}$$

peut être décomposé en termes de variance comportant des lisseurs d'échantillon et de population et des estimateurs paramétriques. Il est possible de démontrer que chacun de ces termes est $o_p(1/n)$. Nous démontrons cette approche pour l'un des termes :

$$\begin{aligned} &\left| \frac{1}{N^2} \sum_{k,l} \sum_{i=1}^n \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l} \frac{\pi_k}{\gamma_k - g_k} \frac{\pi_l}{\gamma_l - g_l} I_k I_l (\mathbf{B} - \mathbf{B}) \right| \\ &\leq \left(C_1 + C_2 \max |\pi_k - \pi_k \pi_l| \right) \frac{1}{n} \sum_{i=1}^n |\bar{\mathbf{z}}_{[s]_1}^k - \bar{\mathbf{z}}_{[s]_1}^k| |\mathbf{B} - \mathbf{B}| \\ &= o_p \left(\frac{n}{1} \right) \end{aligned}$$

où $C_1, C_2 < \infty$ résume les termes bornés (par les hypothèses A1 à A6 et le lemme 1(b)), et le taux de convergence et le résultat de l'hypothèse A6 et des lemmes 1(a) et 2.

Preuve du théorème 3.1 : Dans le lemme 3, nous montrons que

$$\hat{\mathbf{y}}^{\text{reg}} = \hat{\mathbf{y}}_{\text{dir}}^p + o_p \left(\frac{\sqrt{n}}{1} \right),$$

où $\hat{\mathbf{y}}_{\text{dir}}^p$ est l'estimateur par la différence (3). Le résultat découle immédiatement de l'hypothèse A5 et du lemme 4.

(b) les $s^T_K Y^U$ et $s^{UK}_K Z^U$ sont bornées uniformément sur tout $k \in U$.

Preuve du lemme 1 : Puisque les y_k et z_k sont les unes

et les autres bornées par hypothèse, la partie (a) peut être

lemme 4 de Breidt et Opsomer (2000). Bien que ce lemme n'inclue pas de taux de convergence, ce taux est calculé

facilement en notant que

$$\frac{1}{n} \sum_{k \in U} z_k^2 = O\left(\frac{1}{nh}\right)$$

dans la notation de Breidt et Opsomer (2000), puis en

La partie (b) a été prouvée directement dans le lemme 2 (iv) de Breidt et Opsomer (2000).

Lemme 2 Sous les hypothèses A1 à A8,

$$B = B + O_p(1/\sqrt{nh}),$$

avec le taux vérifié en ce qui concerne les composantes, et

B est borné pour tout N .

Preuve du lemme 2 : Écrivons $y^{[s_v]}_k = s^{TK}_K Y^U$ et $y^{[s_v]}_k =$

$s^{AK}_K Y^A$ pour les versions lissées en population et en échantillon de y_k , et, similairement, $z^{[s_v]}_k = s^{UK}_K Z^U$ et

$z^{[s_v]}_k = s^{AK}_K Z^A$. Nous récrivons l'expression (6) sous forme d'une fonction des termes pondérés selon l'échantillon

$$t_l, l = 1, \dots, 6 :$$

$$B = \begin{bmatrix} t_1 & t_2 & t_3 \\ t_2 & t_4 & t_5 \\ t_3 & t_5 & t_6 \end{bmatrix} = \begin{bmatrix} t_1 & t_2 & t_3 \\ t_2 & t_4 & t_5 \\ t_3 & t_5 & t_6 \end{bmatrix}^{-1} \begin{bmatrix} t_6 \\ t_5 \\ t_4 \end{bmatrix},$$

où

$$t_1 = \frac{1}{N} \sum_{k=1}^N z_k^2$$

$$t_2 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k \left(1 - \frac{N}{n}\right)$$

$$t_3 = \frac{1}{N} \sum_{k=1}^N z_k^2 \left(\frac{N}{n}\right)$$

$$t_4 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k + \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k$$

$$t_5 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k \left(1 - \frac{N}{n}\right)$$

$$t_6 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k + \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k.$$

L'estimateur pondéré selon l'échantillon B sera étendu autour de

devons montrer que

$$y^{\text{reg}}_k = y^{\text{dir}}_k + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Lemme 3 Sous les hypothèses A1 à A8, nous avons

A8, du lemme 1(b) et de la limitabilité des z_k .

La limitabilité de B découle directement de l'hypothèse

en appliquant le même raisonnement aux termes restants

de l'inégalité de Markov. Le résultat souhaité s'ensuit alors

deuxième terme est $O_p(1/\sqrt{nh})$ en vertu du lemme 1(a) et

composantes. Le premier terme est borné par A2 et A6, et le

où $z^{[2]}_k$ dénote que les carrés sont calculés pour les

$$\left| \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k \right| \leq \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k \sqrt{\frac{1}{N} \sum_{k=1}^N (y^{[s_v]}_k - y^{[s_v]}_k)^2},$$

terme, nous utilisons l'inégalité de Schwarz

lemme 4 de Breidt et Opsomer (2000). Pour le deuxième

lemme 1(b), en utilisant le même argument que dans le

et le premier terme est $O_p(1/\sqrt{n})$ en vertu de A6 et du

$$+ \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k, \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k$$

$$\left(\frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k \right) \left(\frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - 1 \right)$$

t_6 . Nous avons

démontrons le raisonnement pour l'un de ces termes dans

comportant des quantités lissées $z^{[s_v]}_k$ et $y^{[s_v]}_k$. Nous

A2 et A6. Les termes restants contiennent des sommes

Corollary 5.1.5). Pour t_1 et t_3 , cela découle directement de

$O_p(1/\sqrt{nh})$ pour tout l (par exemple, Fuller (1996),

Taylor d'ordre 0 si nous pouvons montrer que $t_1 - t_1 =$

Le résultat découlera d'un développement en série de

l'inverse en (15), qui est supposée par A8.

lues au point t_1 découle du lemme 1(b) et de l'existence de

et la continuité des dérivées de B par rapport à t_1 et éva-

et les t_l restants peuvent être trouvés dans (15). L'existence

$$t_6 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k + \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k$$

$$t_4 = \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k - \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k + \frac{1}{N} \sum_{k=1}^N z_k^2 y^{[s_v]}_k$$

ou

$$B = \begin{bmatrix} 1 & z^T_N & z^T_N \\ z^T_N & t_4 & t_5 \\ z^T_N & t_5 & t_6 \end{bmatrix}^{-1} \begin{bmatrix} t_6 \\ t_5 \\ t_4 \end{bmatrix}, \quad (15)$$

- **A4** Noyau $K(\cdot)$: le noyau $K(\cdot)$ est symétrique et continu, et satisfait $\int_{-1,1}^1 K(u) du = 1$.
- **A5** Taux d'échantillonnage nN^{-1} et largeur de fenêtre h_n : quand $N \rightarrow \infty$, $nN^{-1} \rightarrow \pi \in (0, 1)$, $h_n \rightarrow 0$ et $Nh_n^2 / (\log \log N) \rightarrow \infty$.
- **A6** Probabilités d'inclusion π_k et π_{kl} : pour tout N , $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k, l \in U_N} \pi_{kl} \geq \lambda' > 0$ et $\limsup_{N \rightarrow \infty} |\pi_{kl} - \pi_k \pi_l| < \infty$.
- **A7** Hypothèses supplémentaires faisant intervenir des probabilités d'inclusion d'ordre supérieur :

$$\lim_{N \rightarrow \infty} n^2 \max_{(k_1, k_2, k_3, k_4) \in D_{k, N}} |E_p(I_{k_1} I_{k_2} I_{k_3} I_{k_4}) (I_{k_3} I_{k_4} - \pi_{k_3 k_4})| < \infty,$$

où $D_{k, N}$ dénote l'ensemble de tous les tuples k distincts (k_1, k_2, \dots, k_l) provenant de U_N .

$$\lim_{N \rightarrow \infty} \max_{(k_1, k_2, k_3, k_4) \in D_{k, N}} |E_p(I_{k_1} I_{k_2} I_{k_3} I_{k_4}) (I_{k_3} I_{k_4} - \pi_{k_3 k_4})| = 0,$$

$$\limsup_{N \rightarrow \infty} \max_{(k_1, k_2, k_3) \in D_{k, N}} |E_p(I_{k_1} I_{k_2} I_{k_3}) (I_{k_2} I_{k_3} - \pi_{k_2 k_3})| < \infty.$$

- **A8** La matrice $N^{-1} Z_U^T (I - S_U^V) Z_U$ est inversible pour tout N avec une probabilité de modèle de 1.

L'hypothèse A8 est requise pour que l'estimateur de population B soit bien défini. L'inversibilité de la matrice en A8 dépend de l'effet combiné de la largeur de fenêtre h et de la distribution conjointe des x_k et z_k . Bien qu'il soit possible, en principe, d'écrire suffisamment de conditions pour cela, nous avons opté pour cette approche plus simple et plus explicite.

Avant de donner les preuves des théorèmes 3.1 et 3.2, nous énonçons et prouvons un certain nombre de lemmes.

Lemme 1 Sous les hypothèses A1 à A7,

- a) pour tout $k \in U$ et $d = 1, \dots, D$,

$$\frac{1}{N} \sum_{d=1}^D E_p(s_{dk}^T Y_d - s_{dk}^T X_d)^2 = O\left(\frac{nh}{1}\right)$$

et

$$\frac{1}{N} \sum_{d=1}^D E_p(s_{dk}^T Z_{da} - s_{dk}^T Z_{du})^2 = O\left(\frac{nh}{1}\right);$$

Remerciements

Il est intéressant de noter que le modèle de régression, il s'agit d'un sujet difficile dans le contexte de l'estimation assistée par un modèle, que complique encore davantage le fait qu'il vient d'être mentionné qu'un seul ensemble de poids de régression sur données d'enquête est appliqué à toutes les variables étudiées : parce que le choix de la largeur de fenêtre optimale dépend de la variable qui est lissée, aucune largeur de fenêtre (et par conséquent aucun ensemble de poids) unique ne sera optimale pour toutes les variables de l'enquête. Ce sujet est étudié à l'heure actuelle par les auteurs.

Hypothèses techniques et calculs

Annexe

Nous commençons par énoncer les hypothèses nécessaires, qui étendent celles utilisées dans Breidt et Opsomer (2000) au modèle semi-paramétrique.

Hypothèses :

- **A1** Distribution des erreurs sous ξ_k : les erreurs ξ_k sont indépendantes et de moyenne nulle, de variance $v(x_k, z_k)$, et à support compact, uniformément pour tout N .

- **A2** Distribution des covariables : les x_k et z_k sont considérées fixes par rapport au modèle de superpopulation ξ_k . Il est supposé que les z_k ont un support borné et que les x_k sont indépendantes et de même loi $F(x) = \int_{-\infty}^x f(t) dt$, où $f(\cdot)$ est une densité avec support compact $[a_x, b_x]$ et $f(x) > 0$ pour tout $x \in [a_x, b_x]$.

- **A3** Fonctions de moyenne et de variance non paramétriques : la fonction de moyenne $m(\cdot)$ est continue et la fonction de variance $v(\cdot, \cdot)$ est bornée et strictement supérieure à 0.

$$F_N^*(i) = \frac{1}{N} \sum_{k \in L} I_{\{y_k \leq i\}}$$

à des valeurs seuil particulières i , où $I_{\{y_k \leq i\}}$ dénote la fonction indicatrice prenant une valeur de 1 si $y_k \leq i$ et 0 autrement. Puisque les trois estimateurs peuvent être exprimés sous forme de sommes pondérées d'observations d'échantillon, les poids obtenus pour chacun peuvent être appliqués directement à la fonction indicatrice $I_{\{y_k \leq i\}}$ pour l'échantillon afin d'estimer $F_N^*(i)$ pour toute valeur i souhaitée. Soit $F_H^*(i)$, $F_{reg}^*(i)$ et $F_{par}^*(i)$ les estimateurs par la régression de Hajek, semi-paramétrique et paramétrique de la fonction de répartition, respectivement. Les estimations de la fonction de répartition de la CNA produites par $F_H^*(i)$, $F_{reg}^*(i)$ et $F_{par}^*(i)$ évaluées sur une grille de 1 000 valeurs uniformément espacées pour i . Sont inclus leurs intervalles de confiance à 95 % ponctuels respectifs calculés en chaque point de la grille. Les trois estimateurs sont comparables, mais les bandes de confiance pour les estimateurs par la régression paramétrique et semi-paramétrique ont tendance à être plus étroites. Si l'on calcule la moyenne sur l'ensemble des 1 000 points de la grille, les largeurs des bandes de confiance sont égales à 0,093 pour $F_H^*(i)$, 0,084 pour $F_{par}^*(i)$ et 0,075 pour $F_{reg}^*(i)$, respectivement.

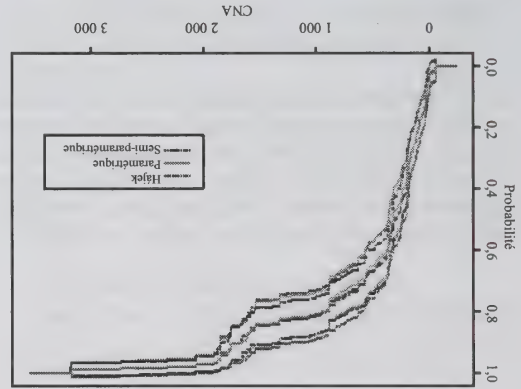


Figure 1
Estimation de la fonction de répartition de population de la CNA et limites de confiance produites par les estimateurs par la régression de Hajek, paramétrique et semi-paramétrique

En plus de la CNA, l'enquête du EMAP sur les lacs du Nord-Est visait à mesurer la concentration de plusieurs substances chimiques, dont les sulfates, le magnésium et les chlorures, de sorte que les poids de sondage obtenus pour la CNA peuvent également être appliqués à ces concentrations, ainsi qu'à leurs fonctions de répartition respectives.

En guise d'illustration supplémentaire de l'approche par estimation semi-paramétrique, il est possible d'« inverser » $F_{reg}^*(i)$ pour obtenir les estimateurs des quantiles $\theta_{reg}^*(\alpha) = \min\{i : F_{reg}^*(i) \geq \alpha\}$ pour ces variables de composition chimique supplémentaires. Le tableau 1 donne les estimations semi-paramétriques des premier, deuxième et troisième quartiles pour les concentrations de sulfates, de magnésium et de chlorures exprimées en $\mu\text{g/L}$. L'estimation de la variance pour ces quantiles pourrait être traitée en utilisant les résultats asymptotiques de Francisco et Fuller (1991), mais nous ne nous pencherons pas sur ce problème ici.

Tableau 1 Estimation des quartiles des variables chimiques

α	Sulfates	Magnésium	Chlorures
0,25	73,3	63,8	27,4
0,50	104,3	127,0	162,2
0,75	201,4	221,9	462,2

5. Conclusion

Dans le présent article, nous avons décrit un estimateur assisté par un modèle qui s'appuie sur la régression semi-paramétrique pour refléter la relation entre de multiples variables auxiliaires au niveau de la population et les variables de sondage. Nous avons élaboré une théorie asymptotique qui montre que l'estimateur résultant est convergent par rapport au plan et asymptotiquement normal sous des hypothèses faibles concernant le plan de sondage et la population. Cette théorie généralise les résultats de Breidt et Opsomer (2000), qui ont prouvé des résultats similaires pour un estimateur assisté par un modèle non paramétrique univarié. L'estimateur semi-paramétrique a été appliqué à des données provenant d'une enquête sur les lacs du Nord-Est des États-Unis, où il s'est avéré plus efficace qu'un estimateur ne tirant pas parti des variables auxiliaires et qu'un estimateur par la régression entièrement paramétrique.

En plus de ses propriétés théoriques, l'estimateur assisté par un modèle semi-paramétrique présente des propriétés pratiques séduisantes. Comme nous l'avons mentionné plus haut, il est entièrement calé pour les variables auxiliaires, qu'elles soient utilisées dans la composante de modélisation paramétrique ou non paramétrique, et il est invariant par rapport à la localisation et à l'échelle. L'estimateur peut être exprimé comme une somme pondérée des observations sur échantillon, de sorte qu'il est conforme aux paradigmes d'estimation par sondage classique et qu'un ensemble unique de poids peut être appliqué à toutes les variables étudiées, ce qui permet de préserver les relations entre les variables.

L'un des problèmes qui n'a pas été abordé dans le présent article est celui de la sélection du paramètre de

Comme sept écorégions différentes sont incluses dans la population, nous construisons des variables muettes $z_{j,k}$ pour $j = 1, \dots, 6$. Nous construisons un estimateur par la régression semi-paramétrique pour la variable y en traitant la variable UTMX x comme un terme non paramétrique et les autres variables $z_1 - z_8$ comme une composante paramétrique. Une approche de sélection de modèle nous a permis de déterminer que le fait de traiter les deux autres variables continues comme étant non paramétriques n'améliorait pas l'ajustement du modèle. Aux fins de comparaison, nous avons également calculé un estimateur par la régression qui traite tous les termes comme étant paramétriques. Cet estimateur est par conséquent identique à l'estimateur semi-paramétrique, excepté que la coordonnée géographique x est modélisée linéairement. Nous noterons cet estimateur par \hat{y}^{par} .

Afin de déterminer l'efficacité estimée des estimateurs d'après les données d'enquête, nous devons calculer les estimations de la variance. Cependant, comme les probabilités d'inclusion de deuxième ordre n'étaient pas disponibles, nous ne pouvons pas évaluer $V(\hat{y}^{\text{reg}})$ comme dans (10). Afin de produire des estimations appropriées de la variance, nous traitons le plan d'échantillonnage complexe comme un plan d'échantillonnage stratifié avec remise. Les 14 strates que nous avons sélectionnées correspondent à des groupes de grappes spatiales de lacs qui figuraient dans le plan d'échantillonnage original et qui ont été utilisées pour assurer la répartition spatiale des lacs échantillonnés sur la région d'intérêt. Larsen et coll. (1993) donnent des précisions sur la construction des grappes spatiales.

Soit H le nombre de strates, n_h le nombre d'observations dans la strate h , et A_h l'ensemble d'éléments échantillonnés qui sont compris dans la strate h . Définissons $p_k = n_h^{-1} \pi_k$. En utilisant cette notation et l'hypothèse d'un échantillon stratifié avec remise, nous récrivons l'estimateur semi-paramétrique sous la forme

$$\hat{y}^{\text{reg}} = \frac{1}{N} \sum_{k \in U} \hat{g}(x_k, z_k) + \frac{1}{N} \sum_{h \in H} \sum_{k \in A_h} \frac{1}{n_h} \frac{p_k}{y_k - \hat{g}(x_k, z_k)} \quad (12)$$

et l'estimateur de la variance sous la forme

$$V(\hat{y}^{\text{reg}}) = \frac{1}{N^2} \sum_{h \in H} \sum_{k \in A_h} \frac{N^2}{S_h^2} \quad (13)$$

où S_h^2 est la variance résiduelle pondérée intrastate estimée pour la strate h . En supposant que les strates sont échantillonnées avec remise, Särndal et coll. (1992, pages 421-422) proposent de calculer S_h^2 comme il suit :

$$S_h^2 = \frac{1}{N^2} \sum_{k \in A_h} \frac{n_h(n_h - 1)}{1} \sum_{l \in A_h} \left(\frac{p_k}{y_k - \hat{g}(x_k, z_k)} - \frac{p_l}{y_l - \hat{g}(x_l, z_l)} \right)^2 \quad (13)$$

De même, nous estimons $V(\hat{y}^H)$ par

$$V(\hat{y}^H) = \frac{1}{N^2} \sum_{h \in H} \frac{n_h(n_h - 1)}{1} \sum_{l \in A_h} \left(\frac{p_k}{y_k - \hat{y}_H} - \frac{p_l}{y_l - \hat{y}_H} \right)^2 \quad (14)$$

et nous obtenons l'expression pour $V(\hat{y}^{\text{par}})$ de façon entièrement analogue à celle utilisée pour $V(\hat{y}^{\text{reg}})$, excepté que $\hat{g}(x_k, z_k)$ est calculé par régression linéaire.

Ces conditions nous permettent d'obtenir les estimations suivantes de la CNA moyenne pour les lacs du Nord-Est, ainsi que les estimations de la variance et les intervalles de confiance (IC) à 95 % approximatif. Nous avons recouru à un ajustement linéaire local pour le terme non paramétrique avec la largeur de la fenêtre fixée à un dixième de l'étendue de l'UTMX.

$$\begin{aligned} \hat{y}^{\text{reg}} &= 558,0 \text{ } \mu\text{eq/L} \quad V(\hat{y}^{\text{reg}}) = 2534,6 \quad \text{IC} = (459,3; 656,6) \\ \hat{y}^{\text{par}} &= 577,3 \text{ } \mu\text{eq/L} \quad V(\hat{y}^{\text{par}}) = 3239,6 \quad \text{IC} = (465,8; 688,9) \\ \hat{y}^H &= 555,9 \text{ } \mu\text{eq/L} \quad V(\hat{y}^H) = 4313,3 \quad \text{IC} = (427,2; 684,7) \end{aligned}$$

de celle de l'estimateur paramétrique.

Comme nous l'avons mentionné plus haut, un objectif important de cette application est de déterminer combien de lacs risqués d'être acidifiés ou le sont déjà. Autrement dit, nous cherchons à estimer la proportion de lacs du Nord-Est dont la valeur de la CNA est inférieure à une valeur seuil particulaire. Nous pouvons déterminer ce genre de proportion en estimant la fonction de répartition en population

S^{bq} , la matrice de lisseurs correspondant pour la variable x_q . En outre, m^{bq} dénote l'estimateur par rétro-ajustement pondéré selon l'échantillon pour $m^q(x_q^{bq})$ et $m^{bq} = (m^{bq}, k \in A)$. L'algorithme de rétro-ajustement pour un modèle contenant des termes non paramétriques \tilde{O} est constitué de l'ensemble d'équations suivant, itéré jusqu'à convergence :

$$\begin{aligned} B &= (Z_T^T \Pi^{-1} Z_T)^{-1} Z_T^T \Pi^{-1} \left(Y - \sum_{q=0}^{b-1} m^{bq} \right) \\ m^{11} &= S^{11} \left(Y - Z_T^T B - \sum_{1 \neq l}^b m^{bq} \right) \\ &\vdots \\ m^{b0} &= S^{b0} \left(Y - Z_T^T B - \sum_{q \neq 0}^b m^{bq} \right). \end{aligned}$$

Ces équations fournissent des ajustements pondérés unique-ment aux points d'observation (localisations) compris dans l'échantillon $k \in A$. Pour les autres points d'observation $k \in U$ non compris dans A , une étape de lissage supplémentaire est nécessaire après l'obtention de $m^{bq}, q = 1, \dots, \tilde{O}$:

$$m^{bq} = S^{bq} \left(Y - Z_T^T B - \sum_{q \neq b}^b m^{bq} \right).$$

Les estimateurs fondés sur l'échantillon de la fonction moyenne à tous les points d'observation $k \in U$ sont alors définis comme étant $\hat{g}_k = m^{11} + \dots + m^{b0} + z_k^b$, qui sont utilisés dans l'expression (8) pour construire l'estimateur assisté par modèle.

4. Application à l'enquête sur les lacs du Nord-Est

À la présente section, nous démontrons l'applicabilité de l'estimateur par la régression semi-paramétrique à un ensemble de données sur la composition chimique d'échantillons d'eau. Comme nous l'illustrerons, une fois que l'on a choisi un ensemble de variables auxiliaires et un modèle, il est aussi facile de calculer des estimateurs pour le modèle semi-paramétrique que pour les modèles linéaires, ce qui peut donc aboutir à une amélioration de la précision à relativement peu de frais.

La National Surface Water Survey (NSWS) paraincée par l'Environmental Protection Agency (EPA) des États-Unis entre 1984 et 1996 a permis d'estimer que 4,2 % des lacs de la région du Nord-Est des États-Unis étaient acides (Stoddard, Kahl, Deviney, DeWalle, Driscoll, Herlihy, Kellogg, Murdoch, Webb et Webster 2003). Les lacs du Nord-Est sensibles aux acides faisaient partie des préoccupations auxquelles visait à répondre le Clean Air Act

Amendement (CAAA) de 1990, grâce à la restriction des émissions industrielles de soufre et d'azote en vue de réduire l'acidité de ces eaux. Une mesure courante de l'acidité est la capacité de neutralisation des acides (CNA), qui est définie comme étant le pouvoir tampon de l'eau, c'est-à-dire la capacité de l'eau de résister aux variations de l'acidité. Une valeur de la CNA inférieure à zéro $\mu\text{eq/L}$ indique que l'eau a perdu son pouvoir tampon. Les eaux de surface dont la valeur de la CNA est inférieure à 200 $\mu\text{eq/L}$ sont considérées comme couvrant un risque d'acidification, et les valeurs inférieures à 50 $\mu\text{eq/L}$ sont considérées comme posant un risque élevé (National Acid Precipitation Assessment Program (1991), page 15).

Entre 1991 et 1996, l'Environmental Monitoring and Assessment Program (EMAP) de l'Environmental Protection Agency des États-Unis a réalisé une enquête sur lacs des États du Nord-Est des États-Unis. Les données ont été recueillies afin de déterminer l'effet que des restrictions imposées par le CAAA ont eu sur les conditions écologiques de ces eaux. Parmi une population de 21 026 lacs, on en a sélectionné pour l'enquête 334, dont certains ont été visités plusieurs fois durant la période d'étude. On a calculé la moyenne des mesures multiples faites sur un même lac afin d'obtenir une seule mesure par lac échantillonné. Les lacs étudiés ont été sélectionnés selon un plan à grille hexagonale utilisée fréquemment par l'EMAP (voir Larsen, Thornton, Urquhart et Paulsen (1993) pour une description du plan d'échantillonnage). Soit y_k la valeur (éventuellement moyenne) de la CNA du k^{e} lac échantillonné. Une estimation très simple de la moyenne des CNA des lacs est représentée par l'estimateur \bar{y}_n . Ici, comme dans le cas de nombreuses enquêtes, un meilleur choix est l'estimateur de Häjek,

$$\hat{y}_H = \frac{1}{N} \sum_{k \in A} \frac{y_k}{\pi_k}, \quad (11)$$

qui applique un ajustement de type ratio pour l'estimation de la taille de population à l'aide de $N = \sum_{k \in A} 1/\pi_k$. Cependant, des variables auxiliaires sont disponibles pour chaque lac dans cette population, si bien qu'il devrait être possible d'améliorer davantage l'efficacité de l'estimateur de Häjek. Les variables qui suivent sont disponibles pour chaque lac $k \in U$:

x_k = UTMX, coordonnée géographique x du centroïde de chaque lac dans le système de coordonnées UTM,
 $z_{j,k}$ = variable indicatrice pour l'écocorégion $j = 1, \dots, 6$,
 $z_{7,k}$ = UTM_Y, coordonnée géographique y ;
 $z_{8,k}$ = élévation.

estimateur à également certaines propriétés souhaitables en commun avec les estimateurs par la régression entièrement paramétrique. Il est invariant par rapport à l'échelle et à la localisation, et il est calé pour les composantes de modélisation paramétriques ainsi que non paramétriques, en ce sens que $\hat{x}_{\text{reg}} = \bar{x}_N$ et $\hat{z}_{\text{reg}} = \bar{z}_N$. Le calage pour les variables figurant dans le terme paramétrique peut être vérifié directement en utilisant les expressions (6) et (7), tandis que le calage pour la variable spécifiée non paramétriquement x_i , découle du fait que $s_{x_i}^T X_i = x_i$, où $X_i = (x_k : k \in A^T)$ (nous ignorons l'effet de l'ajustement diag(δ_N^{-2}) dans (5), parce que ce dernier peut être rendu arbitrairement petit). En outre, l'estimateur peut s'écrire comme une somme pondérée des y_k , $k \in A$, de sorte qu'il est possible d'obtenir et d'appliquer un ensemble de poids w_k à n'importe quelle variable d'enquête d'intérêt.

3. Propriétés et extensions

3.1 Propriétés relatives au plan de sondage

Dans la présente section, nous explorerons les propriétés relatives au plan de sondage de l'estimateur semi-paramétrique (8). Plus précisément, nous prouvons que \hat{y}_{reg} est convergent par rapport au plan au taux \sqrt{n} et nous dérivons sa loi asymptotique, y compris une variance estimée. Cet exercice est fait dans le contexte de convergence asymptotique par rapport au plan utilisé dans Isaki et Fuller (1982) et dans Breidt et Opsomer (2000), dans lequel la taille de la population ainsi que celle des échantillons augmentent quand $N \rightarrow \infty$. Toutes les preuves et les hypothèses nécessaires sont données en annexe.

Dans le théorème qui suit, nous prouvons la convergence par rapport au plan de l'estimateur semi-paramétrique. Nous c'est-à-dire le taux habituel pour les estimateurs fondés sur le plan

Théorème 3.1 *Sous les hypothèses A1 à A8, l'estimateur \hat{y}_{reg} donné par (8) est convergent par rapport au plan au taux \sqrt{n} , en ce sens que*

$$\hat{y}_{\text{reg}} = \bar{y}_N + O_p\left(\frac{1}{\sqrt{n}}\right).$$

Le théorème qui suit prouve qu'il existe un théorème de la limite centrale pour \hat{y}_{reg} , s'il existe pour l'estimateur par extension \bar{y}_N .

Théorème 3.2 *Sous les hypothèses A1 à A8, si*

$$\frac{\bar{y}_N - \bar{y}_N^*}{\sqrt{N}} \rightarrow N(0, 1),$$

En pratique, la formulation de l'algorithme de rétroajustement offre un moyen nettement plus efficace et plus simple de calculer l'estimateur semi-paramétrique. Soit $s_{y_{qk}}$ le vecteur de lisseurs d'échantillon, tel qu'il est défini en (5), pour la variable x_q au point d'observation x_{qk} et

où les $m_q(\cdot)$, $q = 1, \dots, Q$ et $v(\cdot)$, sont des fonctions lisses inconnues.

$$\begin{aligned} E_g(y_k) &= g(x_k, z_k) = m_1(x_k) + \dots + m_Q(x_k) + z_k \beta \\ \text{Var}_g(y_k) &= v(x_k, z_k) \end{aligned}$$

additif semi-paramétrique, qui s'écrit

Les résultats des théorèmes 3.1 et 3.2 sont obtenus en utilisant le modèle semi-paramétrique (1), qui contient un seul terme non paramétrique univarié $m(\cdot)$. Dans de nombreuses applications pratiques, plusieurs variables auxiliaires susceptibles d'être incluses dans la partie non paramétrique du modèle sont disponibles, mais la malédiction de la dimensionnalité fait qu'il est souvent difficile de combiner plusieurs variables en un seul terme non paramétrique multidimensionnel. Au lieu de cela, nous traitons les variables qui doivent être incluses non paramétriquement comme des composantes univariées. Nous obtenons ainsi le modèle

3.2 Modèle additif semi-paramétrique

$$V(\hat{y}_{\text{reg}}) = \frac{1}{N^2} \sum_k \sum_l \frac{\pi_k \pi_l}{\pi_k - \pi_l} \frac{\pi_k}{y_k - \delta_l} \frac{\pi_l}{y_l - \delta_k}. \quad (10)$$

avec

$$\hat{y}_{\text{reg}} = \bar{y}_N + \frac{\sqrt{N}}{\sqrt{N}} \rightarrow N(0, 1),$$

aussi

$$V(\bar{y}_N) = \frac{1}{N^2} \sum_k \sum_l \frac{\pi_k \pi_l}{\pi_k - \pi_l} \frac{\pi_k}{y_k} \frac{\pi_l}{y_l}$$

avec

depend des z_k . Un estimateur plus efficace est obtenu en estimant conjointement $m(\cdot)$ et β , comme le fait l'estimant suivant d'estimateurs

$$\begin{aligned} B &= (Z_T^T(I - S_V^V)Z_V^V)^{-1}Z_T^T(I - S_V^V)Z_V \\ m_k &= s_k^T(K_V - Z_V^V B) \quad k = 1, \dots, N. \end{aligned} \quad (2)$$

Dans ces estimateurs, B est calculé pour commencer, puis le « vecteur résiduel » $K_V - Z_V^V B$ est lissé en fonction de x . Les estimateurs dans (2) sont identiques aux

estimateurs par rétro-ajustement pour les modèles additifs décrits dans Hasle et Tjøsthrant (1990) et implémentés dans gam dans S-Plus, R ou SAS. À titre d'estimateur de population de $E_k(y_k) = g(x|k, z_k)$, nous utilisons

$$g_k = m_k + z_k B.$$

Nous expliquons maintenant comment construire un esti-

mateur semi-paramétrique. Soit $A \subset U$ un échantillon de taille n tiré à partir de U conformément au plan d'échantillonnage $p(A)$ avec les probabilités d'inclusion uni et bidimensionnelles $\pi_k = \sum_{A \ni k} p(A)$, $\pi_{jk} = \sum_{A \ni j, k} p(A)$, respectivement. Si les g_k , $k = 1, \dots, N$ étaient disponibles, il serait possible de construire un estimateur par différence pour la moyenne de population de y_k , $\bar{y}_N = \sum_U y_k / N$ de la forme

$$\bar{y}_{\text{diff}} = \frac{1}{n} \sum_U g_k + \frac{N}{1} \sum_U \frac{\pi_k}{y_k - g_k}, \quad (3)$$

qui est sans biais par rapport au plan et possède une variance due au plan

$$\text{Var}_p(\bar{y}_{\text{diff}}) = \frac{1}{n} \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{\pi_k}{y_k - g_k} \frac{\pi_l}{y_l - g_l}.$$

(Särndal et coll. 1992, page 221). La variance due au plan est faible si les écarts entre y_k et g_k sont petits. Cet estimateur n'est pas faisable, car son calcul nécessite la connaissance de toutes les valeurs de x_k , z_k et y_k pour la population. Nous construisons plutôt un estimateur faisable

en remplaçant les g_k par des estimateurs fondés sur un échantillon. Les estimateurs fondés sur un échantillon correspondent aux estimateurs de population donnés en (2) sont construits de la façon suivante. Le vecteur pondéré selon le plan de lisses polynômiaux locaux est

$$s_{\text{OT}}^{jk} = e^{jT} (X_T^{jk} X^{jk})^{-1} X_T^{jkT} W^{jkT} \quad (4)$$

avec X^{jk} contenant les lignes de X^{jk} qui correspondent à $k \in A$ et

$$W^{jk} = \text{diag} \left\{ \frac{1}{\pi_j h} K \left(\frac{x_j - x_k}{h} \right) : j \in A \right\}.$$

La matrice $X_T^{jk} X^{jk}$ dans (4) sera singulière si, pour un échantillon A , il existe moins que $p + 1$ observations pour appuyer le noyau à un point d'observation x_j . En pratique, ce problème peut être évité en sélectionnant une fenêtre suffisamment large pour rendre la matrice inversible. Cependant, cette situation ne peut pas être écartée en général et nous avons besoin d'un estimateur qui existe pour chaque échantillon A pour les calculs théoriques de la section 3. Donc, nous considérerons le vecteur ajusté de lisses

$$s_T^{jk} = e^{jT} (X_T^{jk} X^{jk} + \text{diag}(\delta N^{-2}))^{-1} X_T^{jkT} W^{jk} \quad (5)$$

pour une distance $\delta > 0$ faible, comme l'ont fait Bredt et Opsomer (2000). La matrice de lisses d'échantillon et sa version centrée sont données par

$$\begin{aligned} S_A^* &= [s_T^{jk} : k \in A] \quad S_A^* = (I - II^T II^T / N) S_A^* \\ B &= (Z_T^T II^T (I - S_V^V) Z_V^V)^{-1} Z_T^T II^T (I - S_V^V) Y_A \end{aligned} \quad (6)$$

$$\hat{m}_k = s_k^T (Y_A - Z_A^V B), \quad (7)$$

où Z_A et Y_A dénotent les versions d'échantillon de Z et Y , respectivement. Notons que l'estimateur \hat{m}_k est défini pour tout point d'observation x_k dans la population, et non pas uniquement pour ceux figurant dans l'échantillon. Comme pour les estimateurs de population, ces estimateurs peuvent s'écrire comme étant la solution d'équations de rétro-ajustement, de sorte qu'ils peuvent être calculés grâce à des versions convenablement pondérées des algorithmes existants. L'estimateur de g_k est

$$\hat{g}_k = \hat{m}_k + z_k B.$$

Nous construisons alors l'estimateur assisté par un modèle semi-paramétrique en remplaçant g_k par \hat{g}_k dans (3) :

$$\bar{y}_{\text{reg}} = \frac{1}{n} \sum_U \hat{g}_k + \frac{1}{N} \sum_U \frac{\pi_k}{y_k - \hat{g}_k}. \quad (8)$$

En définissant $\bar{y}_N^* = \sum_U y_k / \pi_k$ et en faisant de même pour \bar{z}_N^* , nous obtenons pour \bar{y}_{reg}^* une expression équivalente donnée par

$$\bar{y}_{\text{reg}}^* = \bar{y}_N^* + (\bar{z}_N^* - \bar{z}_N^*) B + \frac{1}{N} \sum_U \hat{m}_k - \frac{1}{N} \sum_U \frac{\pi_k}{\hat{m}_k}, \quad (9)$$

qui montre que l'estimateur semi-paramétrique peut être interprété comme étant un estimateur par la régression linéaire « classique » utilisant la composante paramétrique du modèle 2β , avec un terme de correction supplémentaire pour la composante non paramétrique du modèle. Cet

sur échantillon. Cette approche est la même que celle utilisée pour le cas paramétrique dans Samdal et coll. (1992, chapitre 6).

Soit $U = \{1, 2, \dots, N\}$ les étiquettes ordonnées pour une population finie d'intérêt. À titre d'estimateur de population justeement (backfitting estimator) décrit dans Opsomer et Ruppert (1999), Nous commençons par présenter la notation requise. Soit $K(\cdot)$ une fonction noyau utilisée pour définir les voisinages dans lesquels les polynômes locaux seront ajustés (les hypothèses concernant les K sont spécifiées en annexe). Le vecteur de lisses de population pour la régression par polynômes locaux de degré p au point d'observation x_i est défini comme étant

$$s_{iL}^T = e_i^T (X_{iL}^T)^T W_{iL} X_{iL} (X_{iL}^T)^{-1} X_{iL}^T W_{iL}^T$$

où e_i est un vecteur de longueur $p + 1$ avec une valeur 1 à la première position et des valeurs 0 ailleurs, W_{iL} est

$$\text{diag}\{h^{-1}K((x_i - x_1)/h), \dots, h^{-1}K((x_i - x_N)/h)\}$$

$$X_{iL} = \begin{bmatrix} 1 & x_i - x_N & \dots & x_i - x_1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_i - x_N & \dots & x_i - x_1 \end{bmatrix}.$$

Le lisseur s_{iL} peut être appliqué au vecteur $X_{iL}^T = (x_{i1}^{(1)}, \dots, x_{iN}^{(N)})^T$ pour produire l'ajustement de la régression non paramétrique en fonction de la variable x au point d'observation x_i . Il peut aussi être appliqué à n'importe quelle colonne de $Z^T = (z_1^T, \dots, z_N^T)^T$ pour lisser celle-ci par rapport à x . Nous le ferons lors de la dérivation des propriétés de l'estimateur semi-paramétrique (section 3).

En plus du vecteur de lisses au point d'observation x_i , s_{iL}^T , nous devons définir la matrice de lisses à tous les points d'observation x_1, \dots, x_N ,

$$S^U = \begin{bmatrix} s_{1L}^T \\ \vdots \\ s_{NL}^T \end{bmatrix},$$

et la matrice de lisses centrée $S_*^U = (I - 11^T/N)S^U$. Lorsqu'elle est appliquée à X_{iL}^T , la matrice de lisses produit le vecteur des ajustements de la régression non paramétrique à tous les points d'observation. La matrice de lisses centrée S_*^U produit les ajustements centrés, ce qui signifie que la moyenne globale des valeurs ajustées est soustraite de chaque valeur ajustée. Le centrage est utilisé pour préserver l'identifiabilité des estimateurs, comme l'ont expliqué Opsomer et Ruppert (1999).

Pour toute observation x_i , un estimateur possible de $m(x_i)$ pourrait être défini comme étant $s_{iL}^T X_{iL}^T$, avec ou sans ajustement de centrage. Cet estimateur serait généralement médiocre, puisqu'il ne tient pas compte du fait que les y_i contiennent une composante paramétrique qui

Dans Opsomer, Bredt, Moisen et Kauermann (2007), le principe de l'estimation assistée par un modèle non paramétrique a été étendu à des modèles additifs généralisés (MAG) et appliqué dans un modèle d'interaction pour estimer les variables provenant des enquêtes de Forest Inventory and Analysis. Bien que les MAG contiennent aussi un mélange de termes catégoriques (paramétriques) et non paramétriques, ils ne permettent pas d'arriver à un développement théorique complet, si bien que ce dernier n'y a pas été présenté. Le modèle semi-paramétrique faisant l'objet du présent article peut être considéré comme un cas particulier d'un MAG comportant une fonction de lien identifiée. Contrairement au MAG « général », le modèle semi-paramétrique permet la dérivation formelle des propriétés statistiques de l'estimateur assisté par un modèle. La présentation de la suite de l'article est la suivante. À la section 2, nous définissons l'estimateur assisté par un modèle semi-paramétrique. À la section 3, nous énonçons et prouvons les propriétés de l'estimateur par rapport au plan de sondage. À la section 4, nous décrivons l'application de l'estimation assistée par un modèle semi-paramétrique aux données sur les lacs du Nord-Est. Enfin, à la section 5, nous présentons nos conclusions.

2. Estimation assistée par un modèle semi-paramétrique

Pour commencer, nous examinons le modèle de superpopulation contenant un seul terme non paramétrique universel et une composante paramétrique; l'extension à plusieurs termes non paramétriques est abordée à la section 3.2. La composante paramétrique peut être constituée d'un nombre arbitraire de termes linéaires. Il s'agit du modèle semi-paramétrique étudié, entre autre, par Speckman (1988). Ce modèle de superpopulation, que nous désignons par ξ , peut s'écrire sous la forme

$$\begin{aligned} E_{\xi}(y_i) &= g(x_i, z_i) = m(x_i) + z_i \beta \\ \text{Var}_{\xi}(y_i) &= v(x_i, z_i) \end{aligned} \quad (1)$$

où x_i est une variable auxiliaire continue devant être modélisée non paramétriquement et $z_i = (z_{i1}, \dots, z_{iD})$ est un vecteur de D variables auxiliaires catégoriques ou continues qui sont spécifiées paramétriquement. Les fonctions $m(\cdot)$ et $v(\cdot, \cdot)$, ainsi que le vecteur de paramètres β sont inconnus. Pour les besoins d'identifiabilité, nous supposons que le vecteur z_i contient un terme d'ordonnée à l'origine et que la fonction $m(\cdot)$ est centrée autour de 0 en ce qui concerne la distribution de x_i . Nous dériverons l'estimateur assisté par un modèle utilisant le modèle (1) en commençant par définir des estimateurs de population pour les fonctions et paramètres inconnus, puis en construisant des estimateurs

Estimation assistée par un modèle semi-paramétrique pour les enquêtes sur les ressources naturelles

F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson et M. Giovanna Ranalli¹

Résumé

De l'information auxiliaire est souvent utilisée pour améliorer la précision des estimateurs des moyennes et des totaux de la population finie grâce à des techniques d'estimation par le ratio ou par la régression linéaire. Les estimateurs résultants ont de bonnes propriétés théoriques et pratiques, dont l'invariance, le calage et la convergence par rapport au plan de sondage. Cependant, il n'est pas toujours certain que les modèles de ratio et les modèles linéaires sont de bonnes approximations de la relation réelle entre les variables auxiliaires et la variable d'intérêt, ce qui cause une perte d'efficacité si le modèle n'est pas approprié. Dans le présent article, nous expliquons comment on peut étendre l'estimation par la régression afin d'intégrer des modèles de régression semi-paramétriques dans le cas de plans de sondage simples ainsi que plus complexes. Tout en retenant les bonnes propriétés théoriques et pratiques des modèles linéaires, les modèles semi-paramétriques reflètent mieux les relations complexes entre les variables, ce qui se traduit souvent par des gains importants d'efficacité. Nous illustrerons l'applicabilité de l'approche à des plans de sondage complexes comportant de nombreux types de variables auxiliaires en estimant plusieurs caractéristiques liées à l'acidification dans le cas d'une enquête sur les lacs du Nord-Est des États-Unis.

Mots clés : Estimation par la régression; lissage; régression à noyau; chimie des lacs.

1. Introduction

La post-stratification, le calage et l'estimation par la régression sont différentes approches qui permettent d'améliorer la précision des estimateurs lorsqu'on dispose d'information auxiliaire à l'étape de l'estimation. L'*estimation assistée par un modèle* (Samdal, Swensson et Wretman 1992) offre un cadre commode dans lequel élaborer ces estimateurs et des estimateurs connexes. Dans ce cadre, un modèle de superpopulation décrit la relation entre la variable d'intérêt et les variables auxiliaires. Ce modèle est alors utilisé pour construire des estimateurs sur échantillon dont la précision est meilleure si le modèle est spécifié correctement, mais qui retiennent d'importantes propriétés relatives au plan de sondage, comme la convergence et la possibilité d'estimer la variance lorsque le modèle est incorrect. Jusqu'à récemment, les modèles de superpopulation utilisés dans ce contexte étaient formels comme des modèles paramétriques, le plus souvent des modèles de ratio ou des modèles linéaires. Bien que ce genre de modèles relativement simples soient raisonnables dans beaucoup d'applications pratiques, il existe de nombreuses situations où ils ne donnent pas une bonne représentation de la relation entre la variable d'intérêt et les variables auxiliaires. Breidt et Opsomer (2000) ont proposé un estimateur assisté par un modèle non paramétrique basé sur la régression polynomiale locale qui généralise les estimateurs par la régression paramétrique bien établis. Grâce à cet estimateur, la superpopulation ne doit plus nécessairement posséder une forme

paramétrique spécifiée a priori. Au lieu de cela, la relation entre la ou les variables d'intérêt de l'enquête et la variable auxiliaire doit être lisse (continue), mais est par ailleurs entièrement non spécifiée. Dans le présent article, nous étendons en termes formels la théorie de Breidt et Opsomer (2000) au contexte de la régression semi-paramétrique, dans lequel certaines variables sont intégrées linéairement et d'autres le sont à l'aide de termes additifs lisses. Cette extension rend les résultats de ces auteurs plus utiles en pratique, puisque l'information auxiliaire est très fréquemment de nature multidimensionnelle et qu'elle contient presque toujours des variables catégoriques qui doivent entrer paramétriquement dans le modèle de régression (grâce à l'utilisation de variables indicatrices). En guise d'illustration, nous utilisons une enquête sur les lacs des États du Nord-Est des États-Unis réalisées par l'Environmental Monitoring and Assessment Program de l'Environmental Protection Agency des États-Unis. Cette enquête porte sur 334 lacs échantillonnés parmi une population de 21 026 lacs entre 1991 et 1996. Nous appliquons l'estimateur assisté par un modèle semi-paramétrique pour produire des estimations de la moyenne et de la fonction de répartition de la *capacité de neutralisation des acides* et d'autres variables de composition chimique d'intérêt. Dans cette application, nous introduisons linéairement dans le modèle des variables catégoriques ainsi que continues, de même qu'une variable continue à titre de terme additif lisse.

1. F. Jay Breidt, Department of Statistics, Colorado State University, Fort Collins CO 80523, E.-U.; Jean D. Opsomer, Department of Statistics, Iowa State University, Ames, IA 50011, E.-U. Courriel : jopsomer@iastate.edu; Alicia A. Johnson, School of Statistics, University of Minnesota, 224 Church Street SE, Minneapolis MN 55455, E.-U.; M. Giovanna Ranalli, Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via Pascoli, 06123 Perugia, Italie.

- Potter, F. (1990). A study of procedures to identify and trim extreme sample weights. *Proceedings of the Survey Research Section, American Statistical Association*, 1990, 225-230.
- Rizzo, L. (1992). Conditionally consistent estimators using only probabilities of selection in complex sample surveys. *Journal of the American Statistical Association*, 87, 1166-1173.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York : John Wiley & Sons, Inc.
- Särndal, C.-E. (1980). On π -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93, 341-348.
- Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, 40, 364-372.
- Skinner, C.J., Holt, D. et Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York : John Wiley & Sons, Inc.
- Winston, F.K., Kallian, M.K., Elliott, M.R., Menon, R.A. et Durbin, D.R. (2002). Risk of injury to child passengers in compact extended pick-up trucks. *Journal of the American Medical Association*, 287, 1147-1152.

les strates d'inclusion demeurent un important domaine dans lequel poursuivre de futurs travaux.

Remerciements

L'auteur remercie Roderrick J.A. Little, ainsi que le rédacteur en chef, le rédacteur adjoint et deux examinateurs anonymes, de leurs révisions et commentaires. Il remercie aussi les docteurs Dennis Durbin et Flora Winston du projet de la Partners for Child Passenger Safety de leur aide, ainsi que les compagnies d'assurance State Farm de leur appui pour le projet Partners for Child Passenger Safety. La présente étude a été financée par la subvention R01-HL-068987-01 du National Institute of Heart, Lung and Blood.

Bibliographie

- Alexander, C.H., Dahl, S. et Weidman, L. (1997). Making estimates from the American Community Survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 2000, 88-97.
- Association for the Advancement of Automotive Medicine (1990). *The Abbreviated Injury Scale, 1990 Revision*. Association for the Advancement of Automotive Medicine, Des Plaines, Illinois.
- Beaumont, J.-F., et Alavi, A. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue internationale de Statistique*, 51, 279-292.
- Deville, J.-C., et Sarda, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Durbin, D.R., Bhatia, E., Holmes, J.H., Shaw, K.N., Werner, J.V., Sorenson, W., et Winston, F.K. (2001). Partners for child passenger safety: A unique child-specific surveillance system. *Accident Analysis and Prevention*, 33, 407-412.
- Elliott, M.R., et Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, 16, 191-210.
- Ertis, W.A. (1969). Subjective bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society, Series B*, 31, 195-234.
- Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 2000, 598-603.
- Gelfand, A.E., et Smith, A.M.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389-409.
- Gelman, A., et Carlin, J.B. (2002). Poststratification and weighting adjustments. *Survey, Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge et R.J.A. Little), 289-302.
- Gelman, A., et Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- B, 60, 23-40.
- Pfeiffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rabash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.
- Pfeiffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue internationale de Statistique*, 61, 317-337.
- Madow, I. Olkin et D.B. Rubin). 2, 143-184.
- Nonresponse. *Incomplete Data in Sample Surveys*, (Eds., W.G. Oh, H.T., et Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse. *Incomplete Data in Sample Surveys*, (Eds., W.G. McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models*, 2^{ème} édition. CRC Press : Boca Raton, Floride.
- Little, R.J.A. (2004). To model or not model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.
- Little, R.J.A. (1993). Poststratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.
- Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.
- Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.
- Lin, X., et Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61, 381-400.
- Lazzeroni, L.C., et Little, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.
- Lin, X., et Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 65, 175-190.
- Korn, E.L., et Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society, Series B*, 65, 175-190.
- Korn, E.L., et Graubard, B.I. (1999). *Analysis of Health Surveys*. New York : John Wiley & Sons, Inc.
- Korn, E.L., et Graubard, B.I. (1995). Examples of different weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.
- Kish, L. (1992). Weighting for unequal P. *Journal of Official Statistics*, 8, 183-200.
- Isaki, C.T., et Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- Holt, D., et Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, Series A*, 142, 33-46.
- Hastie, T.J., et Tibshirani, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.
- Ghosh, M., et Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.
- A.F.M. Smith). 89-193.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4, Proceedings of the Fourth Valencia International Meeting*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid et

paramètres de régression. Cependant, la spécification parfaite est un objectif impossible à atteindre et, même de bonnes approximations pourraient être fortement biaisées si l'on ne tient pas compte des poids des cas quand les probabilités d'échantillonnage sont très variables. Dans les conditions d'échantillonnage informatif, il pourrait être impossible de déterminer si les divergences entre les estimations pondérées et non pondérées sont dues à une erreur de spécification du modèle ou au plan d'échantillonnage proprement dit. Enfin, même les modèles de régression spécifiques inexactement ont la caractéristique séduisante, dans les conditions de population finie, de produire une seule grandeur de population cible. Par conséquent, nous continuons de recommander de tenir compte de la probabilité d'inclusion sous des modèles linéaires ou des modèles linéaires généralisés, et les méthodes établissant une faible biais et variance élevée et une analyse non pondérée à biais élevé et faible variance demeurent utiles.

Les méthodes dont il est discuté dans le présent article offrent la promesse d'adapter les méthodes fondées sur un modèle pour s'attaquer au problème de l'analyse des données d'enquête. Notre objectif n'est pas d'élaborer un ensemble de données ou à une question particulière, mais plutôt d'élaborer des méthodes robustes, mais efficaces, qui peuvent être appliquées dans les conditions « automatisées » à rythme rapide dans lesquelles sondage appliquée doivent parfois travailler. Quoiqu'elles demandent de nombreux calculs, les méthodes considérées sont des applications ou des extensions de la « boîte à outils » des modèles à effets aléatoires existants et peuvent être implémentées dans les logiciels statistiques existants ou exécutées à l'aide de méthodes MCMC relativement simples. Notre approche conserve un côté fondé sur le plan en ce sens que nous essayons d'élaborer des méthodes d'estimation bayésiennes « automatisées » fondées sur un modèle qui produisent des inférences robustes dans des conditions d'échantillonnage répétées lorsque le modèle proprement dit est spécifié incorrectement. Cependant, parce que ces modèles s'appuient sur la stratification des données par la probabilité de sélection en prélude à l'utilisation des techniques de regroupement en vue de réécarter les données, il existe une correspondance naturelle entre les données, il existe une correspondance naturelle entre cette méthodologie et les plans d'échantillonnage (post) stratifiés dans lesquels les strates correspondent aux probabilités inégales d'inclusion. L'élaboration de méthodes adaptées à une classe plus générale de plans d'échantillonnage complexes comprenant des échantillons en grappes à un ou à plusieurs degrés et (ou) des strates qui « recoupernt »

L'application des méthodes aux données du Partners for Child Passenger Safety en vue de déterminer le risque excédentaire de blessure lors d'un accident chez les enfants installés sur le siège arrière d'une camionnette compacte à cabine allongée comparativement à ceux installés sur le siège arrière d'autres véhicules pour le transport de passagers, il semble que la décision dans Winston et coll. (2002) d'éliminer de l'analyse un enfant dont la probabilité de sélection était faible en vue de stabiliser les estimations était effectivement prudente. En effet, l'estimateur PAA, favorisé par les mesures de l'EQM dans les simulations suggère un risque excédentaire corrigé de 6,7 avec un IFP à 95 % de (3,6, 11,9), comparativement à celui de 14,6 avec un IC à 95 % de (3,4, 61,4) de l'estimateur entièrement pondéré.

Quoique nous appliquions, dans le présent article, une approche entièrement bayésienne de l'inférence au sujet de la loi prédictive de la pente de régression de population, des estimations empiriques bayésiennes (EB) peuvent aussi être obtenues par estimation du maximum de vraisemblance ou du maximum de vraisemblance restreint en utilisant des méthodes standard à modèle mixte linéaire ou linéaire généralisé. Dans les conditions gaussiennes, les estimations EB de G et de σ^2 peuvent être « instables » dans les expressions explicites de $E(B|y, X)$ et $\text{Var}(B|y, X)$. Les conditions exponentielles générales posent plus de problèmes. Les estimations insérées peuvent être utilisées pour déterminer $E(B|y, X)$ par des méthodes de recherche de racine. L'absence de formules explicites pour $E(B|y, X)$ rend difficile l'obtention d'estimateurs empiriques bayésiens basés sur un modèle pour $\text{Var}(B|y, X)$. En outre, les estimateurs empiriques bayésiens standard ne tiennent pas compte de l'incertitude dans l'estimation de G .

Nous notons aussi que, bien que le calcul des valeurs réelles de réduction des poids des cas ne soit pas nécessaire dans cette approche, il est possible de déterminer les poids de sondage révisés qu'implique le rééchantillonnage. Dans les conditions du modèle linéaire, ceux-ci peuvent être obtenus par une application itérative d'un schéma de pondération par calage, tel que les estimateurs par la régression généralisée ou GREG (Deville et Särndal 1992). Les conditions exponentielles générales requièrent l'intégration de l'algorithme de calage des poids dans l'algorithme des moindres carrés itératifs pondérés utilisé pour ajuster un modèle linéaire généralisé.

Lorsque les poids d'échantillonnage sont utilisés pour tenir compte de l'erreur de spécification de la moyenne dans des conditions de régression, on pourrait soutenir que l'approche correcte consiste à spécifier correctement la moyenne afin d'éliminer les divergences entre les estimations entièrement pondérées et non pondérées des

réduction des poids sous forme d'un modèle à effets aléatoires qui lisse les paramètres d'intérêt du modèle sur l'ensemble des classes d'inclusion. Les modèles avec structures de moyenne échangeables offrent le degré le plus important de rétrécissement ou de réduction, mais sont les plus sensibles à l'erreur de spécification du modèle; les modèles dont les moyennes sont fortement structurées pourraient être moins efficaces, mais sont plus robustes à l'erreur de spécification du modèle. Cette propriété de robustesse peut être particulièrement importante, sachant que les éléments des grandes strates d'inclusion donnent le degré le plus important de réduction possible de la variance dans l'estimation fondée sur un modèle, mais sont aussi sujets au degré le plus important de biais et de variance dans le modèle à cause de l'extrapolation.

Nous considérons des simulations sous divers degrés d'erreur de spécification du modèle et d'échantillonnage informatif pour les modèles de régression linéaire et logarithmique. Les modèles de lissage linéaire et non paramétrique surpassent presque les estimateurs entièrement pondérés en ce qui concerne la réduction de l'erreur quadratique dans les simulations considérées. Le modèle d'échangeabilité manifeste une certaine tendance à surlisser, favorisant la réduction de la variance au détriment de la correction du biais, spécialement dans les conditions de régression linéaire. Tous les estimateurs à lissage des poids ont tendance à avoir une couverture des intervalles de confiance inférieure au taux nominal lorsque l'erreur de spécification du modèle est très forte, quoique dans aucun cas, la couverture n'était catastrophiquement faible. Le modèle de lissage local dans les strates de poids semble produire un accroissement non négligeable de l'efficacité tout en posant un risque limité de surlissage ou de sous-dénombrement grave.

Tableau 3
Rapport de cotes estimées d'une blessure chez les enfants assis sur le siège arrière d'une camionnette compacte à cabine allongée ($n = 60$) comparativement à ceux assis sur le siège arrière d'autres véhicules ($n = 8\ 060$), en utilisant les estimateurs non pondérés (NPD), entièrement pondérés (EPD), à (PAL), de la pente aléatoire à modèle d'échangeabilité (PAA), de la pente aléatoire à modèle linéaire (PAL), et de la pente aléatoire à modèle non paramétrique (PANP); non corrigés et corrigés pour l'âge de l'enfant, la gravité de l'accident, la direction de l'impact et le poids du véhicule. Estimations ponctuelles des modèles PAA, PAA et PAL à partir de la médiane à postériori. Intervalle de confiance ou intervalle prédicatif à postériori à 95 % en indice inférieur. Données provenant du Partners for Child Passenger Safety

Non corr.	NPD	EPD	PAL	PAA	PANP
	3,54 (2,01 à 6,23)	11,32 (2,67 à 48,02)	9,15 (2,65 à 31,7)	10,99 (2,97 à 34,64)	10,34 (3,27 à 24,62)
Non corr.	3,50 (1,88 à 6,33)	14,56 (3,45 à 61,40)	10,99 (2,97 à 34,64)	6,70 (2,51 à 20,92)	6,69 (2,64 à 21,05)
Corr.	4,45 (2,39 à 8,67)	6,67 (3,56 à 11,94)	11,87 (3,33 à 36,93)	10,23 (3,02 à 37,93)	10,34 (3,27 à 24,62)

corrigé. Dans les résultats non corrigés, les estimateurs PAA et PAA sont compris entre l'estimateur non pondéré et l'estimateur entièrement pondéré, tandis que les estimateurs linéaires et non paramétriques tendent à suivre l'estimateur entièrement pondéré. Dans l'analyse corrigée, les trois estimateurs fondés sur un modèle sont compris entre l'estimateur non pondéré et l'estimateur entièrement pondéré, l'estimateur PAA étant le plus proche de l'estimateur non pondéré et l'estimateur PAL, le plus proche de l'estimateur entièrement pondéré. Si l'on s'en tient aux résultats des simulations, il semble que l'estimateur PAA, qui suggère des risques relatifs de blessure de l'ordre de 7 pour les enfants passagers dans une camionnette compacte avec cabine allongée comparativement à d'autres véhicules, pourrait être un meilleur estimateur du risque relatif que l'estimateur non pondéré ou entièrement pondéré. (À titre d'évaluation, nous notons que des données portant sur deux années supplémentaires, qui n'étaient pas disponibles au moment de Winston et coll. (2002), englobant 4 091 enfants supplémentaires assis sur le siège arrière de véhicules de transport de passagers [44 dans des camionnettes compactes à cabine allongée] ont donné un rapport de cotes non corrigés entièrement pondérés pour les blessures chez les enfants passagers dans une camionnette compacte à cabine allongée de 6,3, et un rapport de cotes corrigés de 7,0.)

5. Discussion

Les modèles dont il est discuté dans le présent article généralisent les travaux de Lazzeroni et Little (1998), ainsi que d'Elliot et Little (2000) dans lesquels l'inférence à la population a été limitée aux moyennes de population sous des hypothèses de loi gaussienne. Considérer la pondération comme une interaction entre les probabilités d'inclusion et les paramètres du modèle offre un autre paradigme pour la

4. Application : estimation de la prévalence des blessures chez les enfants passagers dans les camionnettes compactes à cabine allongée

L'ensemble de données de Partners for Child Passenger Safety correspond à l'échantillon disproportionné, à probabilités communes, provenant de toutes les décennies 1998 comportant au moins un enfant de 15 ans ou moins passager dans un véhicule de modèle 1990 ou plus récent assuré par State Farm (Durbin, Bhatia, Holmes, Shaw, Werner, Sorenson et Winston 2001). Comme les blessures, et particulièrement les blessures « graves » définies comme étant des lésations faciales ou d'autres blessures recevant une cote de 2 ou plus sur l'échelle AIS (Abbreviated Injury Scale) (Association for the Advancement of Automotive Medicine 1990), sont relativement rares, même chez les enfants dans la population de demandes de remboursement pour dommage à un véhicule accidenté, un échantillon en grappes disproportionnel stratifié est utilisé pour sélectionner les véhicules (l'unité d'échantillonnage) en vue de réaliser un sondage téléphonique auprès des conducteurs. Les véhicules contenant des enfants ayant reçu un traitement médical après l'accident ont été sur-échantillonnés afin que la majorité des enfants blessés soient sélectionnés, tout en veillant à ce que l'échantillon demeure représentatif de l'ensemble de la population. (Par traitement médical, on entend un traitement produit par des ambulanciers paramédiques, un traitement reçu au cabinet d'un médecin ou dans un service d'urgence, ou l'hospitalisation). Pour tout véhicule échantillonné, tous les enfants passagers de ce véhicule ont été inclus dans l'enquête. Les conducteurs des véhicules échantillonnés ont été contactés par téléphone et, si un traitement médical avait été reçu par un passager, soumis à une présélection au moyen d'un questionnaire abrégé afin de confirmer la présence d'au moins un enfant passager ayant subi une blessure. Tous les véhicules contenant au moins un enfant pour lesquels la présélection était positive pour une blessure, ainsi qu'un échantillon aléatoire de 10 % des véhicules pour lesquels il avait été déclaré que les enfants passagers avaient reçu un traitement médical, mais avait donné un résultat de présélection négatif pour les blessures ont été sélectionnés pour un interview complet; un échantillon à 2 % (par après 2,5 %) d'accidents pour lesquels aucun traitement médical n'avait été reçu à également été sélectionné. Parce que la stratification du traitement est imparfaitement associée au risque de blessure (plus de 15 % de la population ayant subi des blessures graves se situe, selon les estimations, dans la catégorie des probabilités de sélection les plus faibles et près de 20 % des personnes n'ayant pas subi de blessure grave sont comprises dans la catégorie des probabilités de

L'ensemble de données des rapports de blessure grave de 2002 ont déterminé que les enfants assis sur le siège arrière dans les camionnettes compactes à cabine allongée courent un plus grand risque de blessure grave que ceux assis sur le siège arrière d'autres véhicules. Cependant, la quantification du degré de risque excédentaire, donc de l'importance du problème de santé publique, posait des difficultés. Le rapport de cotes (RC) non pondérées exprimant le risque d'une blessure grave chez les enfants voyageant dans une camionnette compacte à cabine allongée comparativement à d'autres véhicules était de 3,54 (IC à 95 % : 2,01, 6,23), comparativement à 11,32 (IC à 95 % : 2,67, 48,03) pour blessure ainsi que l'utilisation d'une camionnette compacte à cabine allongée étaient tous deux associés à l'âge de l'enfant, à la gravité de l'accident (intrusion et le fait que le véhicule soit utilisable ou non), à la direction de l'impact et au poids du véhicule, nous avons également considéré un modèle de régression logistique multivariée tenant compte de ces facteurs. Les rapports de cotes corrigés non pondérés et entièrement pondérés pour le risque de blessure chez les enfants assis sur le siège arrière d'une camionnette compacte à cabine allongée comparativement à d'autres véhicules sont de 3,50 (IC à 95 % : 1,88, 6,53) et de 14,56 (IC à 95 % : 3,45, 61,40) respectivement. L'utilisation de l'estimateur non pondéré posait des problèmes, à cause du biais vers zéro induit par le plan d'échantillonnage informatif; toutefois, l'estimateur entièrement pondéré semblait hautement instable, en partie à cause de la présence d'un enfant ayant subi une blessure grave dans une camionnette compacte à cabine allongée pour lequel la probabilité de sélection était très faible (0,025). Dans Winston, Kallian, Elliott, Memon et Durbin (2002) ont normalisée supérieure à 3.

variables : $1 \leq w_i \leq 50$, où 9 % des poids ont une valeur inférieure à cet ensemble de données sont plutôt vers zéro (Korn et Graubard 1995). En outre, les poids qui sélection les plus élevés), le plan d'échantillonnage est informatif, avec des rapports de cotes non pondérés biaisés vers zéro (Korn et Graubard 1995). En outre, les poids qui

Winston, Kallian, Elliott, Memon et Durbin (2002) ont normalisée supérieure à 3.

variables : $1 \leq w_i \leq 50$, où 9 % des poids ont une valeur inférieure à cet ensemble de données sont plutôt vers zéro (Korn et Graubard 1995). En outre, les poids qui

sélection les plus élevés), le plan d'échantillonnage est informatif, avec des rapports de cotes non pondérés biaisés vers zéro (Korn et Graubard 1995). En outre, les poids qui

déterminé que les enfants assis sur le siège arrière dans les camionnettes compactes à cabine allongée courent un plus grand risque de blessure grave que ceux assis sur le siège arrière d'autres véhicules. Cependant, la quantification du degré de risque excédentaire, donc de l'importance du problème de santé publique, posait des difficultés. Le rapport de cotes (RC) non pondérées exprimant le risque d'une blessure grave chez les enfants voyageant dans une camionnette compacte à cabine allongée comparativement à d'autres véhicules était de 3,54 (IC à 95 % : 2,01, 6,23), comparativement à 11,32 (IC à 95 % : 2,67, 48,03) pour blessure ainsi que l'utilisation d'une camionnette compacte à cabine allongée étaient tous deux associés à l'âge de l'enfant, à la gravité de l'accident (intrusion et le fait que le véhicule soit utilisable ou non), à la direction de l'impact et au poids du véhicule, nous avons également considéré un modèle de régression logistique multivariée tenant compte de ces facteurs. Les rapports de cotes corrigés non pondérés et entièrement pondérés pour le risque de blessure chez les enfants assis sur le siège arrière d'une camionnette compacte à cabine allongée comparativement à d'autres véhicules sont de 3,50 (IC à 95 % : 1,88, 6,53) et de 14,56 (IC à 95 % : 3,45, 61,40) respectivement. L'utilisation de l'estimateur non pondéré posait des problèmes, à cause du biais vers zéro induit par le plan d'échantillonnage informatif; toutefois, l'estimateur entièrement pondéré semblait hautement instable, en partie à cause de la présence d'un enfant ayant subi une blessure grave dans une camionnette compacte à cabine allongée pour lequel la probabilité de sélection était très faible (0,025). Dans Winston et coll. (2002), cet enfant a été éliminé avant d'effectuer l'analyse.

Le tableau 3 donne les résultats pour les rapports de cotes non corrigés et corrigés du risque de blessure grave obtenus en utilisant les estimateurs fondés sur le plan de sondage non pondéré, entièrement pondéré et à poids réduits, ainsi que les estimateurs fondés sur les modèles de la pente de la régression à échangeabilité, autorégressif et linéaire. (Résultats pour les estimateurs basés sur un modèle d'après 250 000 tirages d'une seule chaîne après un tirage de rodage de 50 000; la convergence a été évaluée par la méthode de Geweke (1992).) Pour les modèles PAE et PAA, $p(\Sigma) \sim \text{INVERSE-WISHART}(p, 0,11)$, où $p = 2$ pour le modèle non corrigé et $p = 13$ pour le modèle

La pente linéaire est approximativement correctement spécifiée, mais donne de mauvais résultats pour un degré modéré à important d'erreur de spécification. Les estimateurs à poids réduits, autorégressifs et non paramétrique sont tous supérieurs à l'estimateur standard entièrement pondéré, et l'estimateur à échangeabilité et l'estimateur linéaire le sont presque, sur l'étendue des simulations considérées. L'estimateur à poids réduits brnt a produit une réduction de 30 % de l'EQM, les estimateurs non paramétriques, à échantillon et autorégressifs, des réductions allant jusqu'à 20 %

25 %, et l'estimateur linéaire, des réductions de seulement 10 % ou moins.

L'estimateur non pondéré donne une mauvaise couverture de 1C nominal, sauf quand le modèle à pente linéaire est spécifié correctement ou qu'il est presque. Les estimateurs fondés sur un modèle ont généralement de bonnes propriétés de couverture lorsque le modèle linéaire a été spécifié correctement, une légère réduction de la couverture étant observée lorsque la courbe est importante.

Tableau 1
Biais relatif (%), racine carrée de l'erreur quadratique moyenne (REQM) relativement à la REQM de l'estimateur entièrement pondéré, et couverture réelle de l'intervalle de confiance ou de l'intervalle prédictif à posteriori à 95 % de l'estimateur de la pente de la régression linéaire de population sous erreur de spécification du modèle. La pente et l'ordonnée à l'origine de population sont estimées au moyen des estimateurs fondés sur le plan de sondage non pondéré (NPD), entièrement pondéré (EPD), et à poids réduits (PAA), et en tant que moyenne à posteriori dans (8) sous une loi a priori d'échangeabilité (PAB), autorégressif (PAA), linéaire (PAL) et non paramétrique (PANP) pour les paramètres de régression. Les valeurs de l'EQM relatives à l'estimateur entièrement pondéré inférieures à 1 sont en caractères gras

Estimateur	Biais relatif (%)					log ₁₀ de la variance					log ₁₀ de la variance					log ₁₀ de la variance					log ₁₀ de la variance				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
NPD	21,5	21,8	22,2	20,8	22,3	12,1	4,57	1,76	0,75	0,67	0	0	6	78	92	0	0	6	78	92	0	0,2	0,04	0,06	0,08
EPD	0,0	0,1	1,4	1,6	-0,2	4,74	1,88	1,02	0,71	0,75	94	95	96	94	96	94	95	96	94	96	94	95	96	94	96
PEL	8,3	8,4	9,6	8,8	7,8	4,74	1,88	1,02	0,71	0,75	94	95	96	94	96	94	95	96	94	96	94	95	96	94	96
PAA	0,1	1,4	9,6	14,5	17,4	1,00	1,03	1,11	0,74	0,69	87	89	78	90	96	87	89	78	90	96	87	89	78	90	96
PAL	-0,2	-0,4	1,1	1,6	-0,3	0,99	0,91	0,91	0,91	0,93	85	91	96	95	94	85	91	96	95	94	85	91	96	95	94
PANP	-0,1	-0,3	0,9	1,5	-0,4	0,89	0,90	0,95	0,90	0,95	86	92	96	94	94	86	92	96	94	94	86	92	96	94	94

Tableau 2
Biais relatif (%), racine carrée de l'erreur quadratique moyenne (REQM) relativement à la REQM de l'estimateur entièrement pondéré, et couverture réelle de l'intervalle de confiance ou de l'intervalle prédictif à posteriori à 95 % de l'estimateur de la pente de la régression logistique de population sous erreur de spécification du modèle. La pente et l'ordonnée à l'origine de population sont estimées au moyen des estimateurs fondés sur le plan de sondage non pondéré (NPD), entièrement pondéré (EPD) et à poids réduits (PEL), et en tant que moyenne à posteriori dans (8) sous une loi a priori d'échangeabilité (PAB), autorégressif (PAA), linéaire (PAL) et non paramétrique (PANP) pour les paramètres de régression. Les valeurs de l'EQM relatives à l'estimateur entièrement pondéré inférieures à 1 sont en caractères gras

Estimateur	Biais relatif (%)					log ₁₀ de la variance					log ₁₀ de la variance					log ₁₀ de la variance					log ₁₀ de la variance				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
NPD	4,9	-11,9	-21,6	-34,6	0,57	0,73	0,88	1,19	1,61	0,90	95	96	89	66	32	17	0,02	0,04	0,06	0,08	0,02	0,04	0,06	0,08	0,02
EPD	1,1	2,2	1,3	-0,3	1,6	1	1	1	1	1	95	94	90	94	94	94	95	94	90	94	95	94	90	94	94
PEL	0,5	-1,0	-3,5	-7,2	-12,1	0,70	0,77	0,77	0,78	0,95	98	97	94	84	92	92	98	97	94	84	92	97	94	84	92
PAA	1,3	-0,8	-1,9	-5,6	-8,7	0,75	0,82	0,85	0,88	1,02	97	97	94	92	91	90	97	97	94	92	97	97	94	92	91
PAL	0,8	1,7	1,5	-0,4	1,1	0,89	0,97	0,94	0,91	1,02	95	91	88	92	92	89	95	91	88	92	95	91	88	92	92
PANP	0,3	1,5	1,1	0,9	0,5	0,87	0,88	0,87	0,80	0,90	95	92	88	94	96	96	95	92	88	94	96	92	88	94	96

trois estimateurs fondés sur le plan de sondage et les quatre estimateurs fondés sur un modèle de la pente de la population (β) étudiée, en fonction de la variance σ^2 .

L'estimateur entièrement pondéré de la pente de la population est essentiellement sans biais par rapport au plan sous erreur de spécification du modèle, les estimateurs non pondérés et à poids réduits sont biaisés. Les biais des modèles d'échangeabilité et autorégressifs augmentent à mesure que la variance augmente, car ces modèles changent l'absence de biais de l'estimateur entièrement pondéré pour la variance réduite de l'estimateur non pondéré. Les modèles linéaire et non paramétrique sont approximativement

sans biais. Les estimateurs non pondérés et à poids réduits donnent

de mauvais résultats en ce qui concerne l'EQM pour les petites valeurs de σ^2 , où le biais dû à l'erreur de spécification du modèle est critique, et de bons résultats pour les valeurs plus grandes de σ^2 , où l'instabilité de l'estimateur entièrement pondéré est plus importante que la réduction du biais. L'estimateur basé sur un modèle d'échangeabilité a de bonnes propriétés en ce qui concerne la REQM pour les petites et les grandes valeurs de σ^2 , les réductions de l'EQM étant de plus de 30 %, mais produit un message excessif pour les degrés intermédiaires de spécification du modèle. Les propriétés du modèle autorégressif sont égales

à celles du modèle d'échangeabilité pour les petites et les grandes valeurs de σ^2 , mais il est protégé en grande partie contre le surlissage observé pour le modèle d'échangeabilité aux niveaux intermédiaires. Les modèles linéaire et non paramétrique dominent essentiellement les estimateurs entièrement pondérés en ce qui concerne l'EOM sous toutes les simulations envisagées, quoique les réductions de l'EOM soient seulement de l'ordre de 10 %.

mauvaise couverture de l'intervalle de confiance, sauf quand l'erreur de spécification du modèle est quasi absente. L'échec du compromis entre le biais et la variance dans le cas de l'estimateur d'échangeabilité en présence d'erreur de spécification du modèle est mis en évidence par la mauvaise couverture de l'estimateur pour les valeurs intermédiaires de σ^2 ; cette situation est améliorée, mais n'est pas entièrement éliminée, dans le cas de l'estimateur autorégressif. Les estimateurs linéaire et non paramétrique ont une bonne couverture lorsque l'erreur de spécification du modèle est moins importante, mais produisent une certaine sous-couverture lorsque l'erreur de spécification du modèle est plus importante.

3.2 Régression logistique

Dans le modèle de régression logistique, nous avons généré des données de population de la façon suivante :

$$P(Y^i = 1 | X^i) \sim B(\expit(3.25 - 0.75X^i + \gamma X^i_2)), \quad (12)$$

$$X_i \sim U(0, 1), i = 1, \dots, N = 20000$$

où $B(p)$ est une loi de Bernoulli avec probabilité de « succès » p , $\exp(-) = \exp(-)/(1 + \exp(-))$. L'objet de l'analyse est d'obtenir la pente de population de la régression logistique définie comme étant la valeur B , dans l'équation $\sum_{i=1}^n (y_i - \expit(B_0 + B \cdot x_i)) = 0$. Un plan d'échantillonnage avec probabilités de sélection intégrales a été appliqué comme il est décrit dans les simulations par la régression linéaire. Nous considérons les valeurs de $\gamma = 0,10158, 0,0273, 0,0368, 0,0454$, qui correspondent à des mesures de courbes de $K = 0, 0,02, 0,04, 0,06, 0,08$ au point médian z du support de X , où $K(\gamma; X) = |z|^{2/\gamma}/(1 + (2\gamma X - 0,075)^{2/\gamma})$. 200 simulations ont été exé-

cutées pour chaque valeur de γ . Comme dans les simulations sans remise proportionnelle, les éléments ont été échantillonnés sans remise avec probabilité proportionnelle à $(1 + h/2,5)^h$; en tout, 1 000 éléments ont été échantillonnés pour chaque simulation. Nous avons de nouveau considéré les estimateurs basés sur l'EPMV entièrement pondéré (EPD), non pondéré (NPD) et à poids réduits (FEL), ainsi que les estimateurs à pente aléatoire échantillonnée (PAL), l'inférence au sujet des estimateurs paramétrique (PANP), l'inférence au moyen d'approximations par développement en série de Taylor (Binder 1983), comme il est discuté à la section précédente.

REQM de l'estimateur entièrement pondéré et la couverture réelle des IC ou des intervalles prédictifs à posteriori à taux nominal de 95 % associée à chacun des sept estimateurs de la pente de population (B) pour diverses valeurs de la courbure K , correspondant à des degrés croissants d'erreur de spécification.

Le sous-échantillonnage des petites valeurs de X

signifie que l'estimateur du maximum de vraisemblance de B dans les conditions "0" est de spécification du modèle est sans biais tout $K = 0$ et présente un biais par défaut pour $K = 0,02, 0,04, 0,06$, et $0,08$ à moins qu'il soit tenu compte du plan d'échantillonnage. Le biais de l'estimateur à poids réduits est compris entre ceux de l'estimateur non pondéré et de l'estimateur entièrement pondéré. Le biais de l'estimateur à échangeabilité est compris entre ceux de l'estimateur entièrement pondéré et de l'estimateur autoregressif est compris entre ceux de l'estimateur à échangeabilité et de l'estimateur

entièrement pondéré, tandis que les estimateurs linéaire et non paramétrique sont essentiellement sans biais. L'estimateur non pondéré possède une EQM sensible-ment améliorée (40 % de réduction) lorsque le modèle à

$$P(I_i = 1 | h_i) = \pi_i \propto (1 + h_i/2,5) h_i$$

Nous avons créé ainsi 10 strates, définies par les parties entières des valeurs de X_i . Les éléments (Y_i, X_i) avaient $\approx 1/36$ de la probabilité de sélection quand $0 < X_i \leq 1$ que lorsque $9 < X_i < 10$. Nous avons échantillonné $n = 500$ éléments sans remise pour chaque simulation. L'objet de l'analyse était d'obtenir la pente de population $B_i = \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^N (X_i - \bar{X})^2$. Nous avons fixé $\alpha = \beta = 1$, ce qui a introduit un biais positif dans l'estimation de B_i , et avons fait varier σ^2 . L'effet de l'erreur de spécification du modèle augmente à mesure que $\sigma^2 \rightarrow 0$, car le biais des estimateurs devient grand comparativement à la variance, et inversement diminue quand $\sigma^2 \rightarrow \infty$. Nous avons considéré les valeurs de $\sigma^2 = 10^0, 10^1, \dots, 5, 200$ simulations ont été générées pour chaque valeur de σ^2 .

Ici et plus loin, nous avons utilisé une hypothèse a priori Wishart inverse sur la variance a priori G_i , centrée à la matrice d'identité avec deux degrés de liberté.

En plus des modèles à pente aléatoire échangeable (PAB), à pente aléatoire autorégressive (PAA), à pente aléatoire linéaire (PAL) et à pente aléatoire non paramétrique (PANP) discutés à la section 2.3, nous avons considéré l'estimateur fondé sur le plan (entièrement pondéré) standard, ainsi que les estimateurs à poids réduits et non pondéré. Pour l'estimateur entièrement pondéré (EPD), nous avons utilisé l'EPMV $B_i = (X_i'W_iX_i)^{-1}X_i'W_iY_i$ où, en dénotant par une minuscule les éléments échantillonnés $(I_i = 1)$, $w_i \equiv w_h$ pour $h = 1, \dots, H$, $i = 1, \dots, n_h$, $W = \text{diag}(w_h)$, $x_h = (1, x_h)'$, X_h contient les lignes emplies de x_h' et X contient les matrices emplies X_h . Nous avons obtenu l'inférence au sujet de B_i par l'approximation standard par développement en série de Taylor (Binder 1983) :

$$\text{Var}(\hat{B}_i) = \hat{S}_{XX}^{-1} \sum_{i=1}^N (\hat{B}_i) \hat{S}_{XX}^{-1}$$

où \hat{S} est un estimateur convergent d'un total de population $\sum_{i=1}^N x_i'x_i$ donné par $X'WX$ et $\sum_{i=1}^N (\hat{B}_i)$ est une estimation convergente sous le plan de la variance entre la valeur de y_i et sa $e_i = y_i - x_i'B$ est la différence entre la valeur de y_i et sa valeur estimée sous la pente de population réelle B : $\sum_{i=1}^N (\hat{B}_i) = \sum_{h=1}^H n_h (1 - \sum_{i=1}^{n_h} (\hat{x}_h - \bar{x}_h)(\hat{x}_h - \bar{x}_h))$, où $\bar{x}_h = y_h / e_h$ pour $e_h = y_h - x_h'B$. Nous considérons aussi l'estimateur à poids réduits (PEL) obtenu en remplaçant les poids w_h par les valeurs réduites w_h' qui fixent la valeur maximale normalisée à 3 : $w_h' = N w_h / \sum_{i=1}^{n_h} w_h$, où $w_h' = \min(w_h, 3N/n_h)$, et l'estimateur non pondéré (NPD) obtenu en fixant $w_h' = N/n$ pour tous h, i .

Le tableau 1 donne le biais relatif, la racine de l'erreur quadratique moyenne (REQM) et la couverture réelle de l'intervalle de confiance à taux nominal de 95 % pour les

Le rétrécissement le plus important, correspondant à la

réduction le plus rigoureux des poids, s'obtient quand la variabilité des strates de poids est faible, ou quand les strates des probabilités d'inclusion les plus faibles sont mal estimées. Le rétrécissement devrait être faible si les pentes de strate de poids sont estimées avec précision et qu'elles sont systématiquement associées à leurs probabilités d'inclusion. En nous basant sur Elliott et Little (2000), nous nous attendrions à ce que le modèle PAB soit le plus efficace lorsqu'une réduction importante des poids est nécessaire pour minimiser l'EQM, mais soit le plus vulnérable au « sur-rétrécissement » quand la correction du biais est l'aspect le plus important. Accroître la structure, particulièrement dans la partie moyenne du modèle, comme dans les modèles PAL et PANP, donnera une estimation plus robuste en ce sens que le sur-rétrécissement aura lieu uniquement dans les situations quasi pathologiques (par exemple, lorsque les tendances des moyennes sont non monotones et très discontinues) et, même dans de telles situations, pourrait donner lieu à une correction du biais un peu plus faible que ne le justifient les données. Le prix à payer pour cette robustesse sera toutefois une réduction de l'efficacité relative des modèles d'échangeabilité.

3. Résultats des simulations

Parce que nous souhaitons des modèles simultanément plus efficaces que les estimateurs fondés sur le plan de sondage, mais néanmoins raisonnablement robustes à l'erreur de spécification du modèle - et en général nous estimons que même les modèles bayésiens devraient avoir de bonnes propriétés fréquentistes - nous évaluons nos modèles proposés dans un contexte de rééchantillonnage. Nous considérons les régressions linéaire et logistique, sous un modèle spécifié incorrectement avec un plan d'échantillonnage non informatif.

3.1 Régression linéaire

Dans le cas du modèle de régression linéaire en présence d'erreur de spécification, nous avons généré des données de population de la façon suivante :

$$X_i | X_i, \sigma^2 \sim N(\alpha X_i + \beta X_i^2, \sigma^2), \quad (11)$$
$$X_i \sim U(0, 10), i = 1, \dots, N = 20\,000.$$

Nous avons utilisé un schéma d'échantillonnage proportionnellement spécifié, non informatif, pour échantillonner les éléments en fonction de X_i (I_i égale 1 si l'élément est échantillonné et 0 autrement) :

$$h_i = \lceil X_i \rceil$$

substitution dans (8), ou β^h si une hypothèse a priori non informative est appliquée à β^h et que sa moyenne a posteriori est obtenue par $(x^T x)^{-1} x^T y$. Les méthodes empiriques ou entièrement bayésiennes qui permettent que les données estiment les paramètres de deuxième degré admettent donc le « lissage des poids » dicté par les données, qui établit un compromis entre les estimateurs non pondéré et entièrement pondéré.

En pratique, naturellement, la moyenne et les composantes de la variance de deuxième degré sont en général inconnues; donc, nous complétons les spécifications du modèle en postulant une hypothèse a priori pour les paramètres de deuxième degré :

$$p(\phi, \beta^h, G) \propto p(G). \quad (5)$$

Habituellement, l'hypothèse a priori $p(G)$ est faiblement informative ou non informative. Nous pouvons alors utiliser l'échantillonnage de Gibbs (Gelfand et Smith 1990; Gelman et Rubin 1992) pour obtenir des tirages à partir de la loi

à posteriori conjointe complète de $(\beta^h, \beta^0, \phi, G)^T | y, X$. Dans le modèle PAE, nous considérons $p(\alpha, \beta^h, \Sigma) \propto \sigma^{-2} \exp(-1/2 \{r^T \Sigma^{-1} r\})$, c'est-à-dire des lois

à priori non informatives pour les paramètres d'échelle et de moyenne a priori, ainsi qu'une hypothèse a priori Wishart-inverse de la variance a priori G centrée à la matrice d'identité à facteur d'échelle r avec p degrés de liberté. La même loi a priori est utilisée pour le modèle PAA, avec l'hypothèse supplémentaire que $p \sim U(0, 1)$ (autocorrélation non négative entre les strates d'inclusion). Dans les modèles PAL et PANP, $p(\alpha, \beta^h, \Lambda) \propto \sigma^{-2}$ et $p(\alpha, \beta^h, \tau) \propto \sigma^{-2}$ (loi a priori non informative du paramètre d'échelle et hyperloi a priori standard). La description des tirages conditionnels de l'échantillonnage de Gibbs peut être consultée à <http://www.sph.umich.edu/mreliot/trim/method2.pdf>.

Le degré de compromis est une fonction de la structure de la moyenne et de la variance du modèle choisi. Les modèles PAE et PAA reposent sur l'hypothèse que les moyennes de pentes sont échangeables; le modèle PAA est plus souple en ce sens que la structure de sa variance permet que les unités ayant des probabilités d'inclusion presque égales soient lissées plus fortement que celles dont les probabilités d'inclusion sont très inégales. Le modèle PAL suppose une tendance linéaire sous-jacente dans les pentes, tandis que le modèle PANP suppose uniquement une tendance sous-jacente lissée jusqu'à la deuxième dérivée. Notons que, dans les modèles PAL et PANP, nous supposons l'indépendance a priori des paramètres de régression associés à une covariable donnée, c'est-à-dire $(\beta_{1j}, \dots, \beta_{Hj}) \perp (\beta_{1j'}, \dots, \beta_{Hj'})$, $j \neq j'$. Il en est ainsi parce que nous modélisons les tendances dans ces paramètres sur l'ensemble de la strate d'inclusion et que nous ne souignons pas « lier » ces tendances sur l'ensemble des covariables.

$$\left\{ \int f_j : f_j^i \text{ absolument continue, } v = 0, 1, \int f_{(2)}^j(u)(u)^2 du > 0, \min_j \sum_h (\beta_{hj}^i - f_j^i(h))^2 + \lambda_j \int f_{(2)}^j(u)(u)^2 du \right\}$$

$$\beta^h = (f_0^h(h), f^h(h), \dots, f^p(h)), G = 0.$$

4. Pente aléatoire non paramétrique (PANP) :

$$G = I^h \otimes \Lambda.$$

$$\beta^h = (\beta_{00}^h + \beta_{01}^h, \dots, \beta_{p0}^h + \beta_{p1}^h),$$

3. Pente aléatoire linéaire (PAL) :

où h indice de nouveau la probabilité d'inclusion, I^h est une matrice d'identité $H \times H$, p est un paramètre d'autocorrélation qui contrôle le degré de rétrécissement sur l'ensemble de strates de poids, Σ est une matrice de covariance $p \times p$ non contractile, Λ est une matrice diagonale $p \times p$ et $f_j(h)$ est une fonction lisse doublement dérivable de h qui minimise la somme des carrés des résidus plus une pénalité d'irrégularité paramétrisée par λ_j (Wahba 1978, Hastie et Tibshirani 1990). En reformulant le modèle PANP comme dans Wang (1998) nous obtenons

$$y^h | \beta^h \sim N(x^h T^h \beta^h, \sigma^2) \\ \beta_{hj} = \beta_{j0}^h + \beta_{j1}^h h + \omega^h n_j \\ n_j \sim N^{H-1}(0, I \tau_j^2), \tau_j^2 = \sigma^2 / (H \lambda_j^f) \quad j = 0, \dots, p$$

où ω^h est la i^e ligne de la décomposition de Choleski de la matrice de base des splines cubiques Ω , où $\omega_k^h = \int_0^h ((h-1)/(H-1) + ((h-1)/(H-1) - 1)^+ d, (x)^+ = x$ si $x \geq 0$ et $(x)^+ = 0$ si $x < 0$, $h, k = 1, \dots, H$. Le modèle PANP peut être étendu à la forme du modèle linéaire généralisé comme dans Lin et Zhang (1999), où l'hypothèse de normalité au premier degré est remplacée par une fonction lién qui est linéaire en les covariables : $g(E(y^h | \beta^h)) = x^h T^h \beta^h$ pour $g(\cdot)$ comme dans (4).

En supposant pour le moment que les paramètres de deuxième degré sont connus, nous voyons que, dans le cas du modèle PAE avec des données normales, à mesure que $|G| \rightarrow \infty$, le partage d'information entre les strates d'inclusion cesse et $\beta^h \approx (x^h x^h)^{-1} x^h y^h$, l'estimateur par la régression dans la strate d'inclusion. En introduisant cette expression par substitution dans (8) nous obtenons $\beta \approx \beta^h$, l'estimateur entièrement pondéré de la pente de population. De même, à mesure que $|G| \rightarrow 0$, les pentes dans les strates d'inclusion $\beta^h \approx \beta^0$, qui est la pente a priori commune, ce qui donne $\beta \approx \beta^0$ après introduction par

rétablit l'estimateur entièrement pondéré, comme nous l'avons montré plus haut. Par ailleurs, à mesure que $\tau^2 \rightarrow 0$, $w_h \rightarrow 0$ de sorte que $E(\mu_h | X^{\text{obs}}) \rightarrow \bar{y}^{\tau^2=0} = \bar{y}$, la moyenne non pondérée; donc, les unités exclues de l'échantillon sont estimées à la moyenne regroupée, puisque le modèle suppose que tous les y_{hi} sont tirés à partir d'une moyenne commune. Par conséquent, ce modèle de lissage des poids permet un compromis entre l'estimateur convergeant sous le plan qui pourrait être très inefficace et l'estimateur non pondéré qui est entièrement efficace sous l'hypothèse forte que les probabilités d'inclusion et la moyenne de Y sont indépendantes. En supposant que τ^2 suit une hyperloi à priori faible, le degré de compromis entre les moyennes pondérée et non pondérée sera « dicté par les données », quoique sous les hypothèses de modélisation.

2.3 Lissage des poids pour les modèles de régression linéaire et de régression linéaire généralisée

Les modèles de régression linéaire généralisée (McCullagh et Nelder 1989) postulent pour y_i une vraisemblance de la forme

$$f(y_i; \theta, \phi) = \exp \left[\eta_i y_i - b(\theta_i) + c(y_i; \phi) \right] \quad (4)$$

où $a_i(\phi)$ fait intervenir une constante connue et un paramètre d'échelle (de perturbation) ϕ , et la moyenne de y_i est reliée à une combinaison linéaire de covariables fixes x_i par une fonction lien $g(\cdot) : E(y_i | \theta_i) = \mu_i$, où $g(\mu_i) = g(b'(\theta_i)) = \mu_i = x_i^T \beta$. Nous avons aussi $\text{Var}(y_i | \theta_i) = a_i(\phi) V(\mu_i)$, où $V(\mu_i) = b''(\theta_i)$. Le lien est canonique si $\theta_i = \eta_i$, auquel cas $g(\mu_i) = V^{-1}(\mu_i)$. Des exemples bien connus sont la loi normale, où $a_i(\phi) = \sigma^2$ et le lien canonique est $g(\mu_i) = \mu_i$; la loi binomiale, où $a_i(\phi) = n_i^{-1}$ et le lien canonique est $g(\mu_i) = \log g(\mu_i) = \log(\mu_i/(1 - \mu_i))$; et la loi de Poisson, où $a_i(\phi) = 1$ et le lien canonique est $g(\mu_i) = \log(\mu_i)$.

$$(\beta^T, \dots, \beta^T)^T | \beta^*, G \sim N^{Hh}(\beta^*, G), \quad (5)$$

où β^* est un vecteur inconnu des valeurs moyennes des coefficients de régression et G est une matrice de covariances inconnue. Nous considérons que la quantité de population cible d'intérêt $\mathbf{B} = (B^1, \dots, B^p)^T$ est la pente qui résout l'équation de score de population $U^N(\mathbf{B}) = 0$, où

Notons que la quantité \mathbf{B} telle que $U(\mathbf{B}) = 0$ est toujours une quantité de population d'intérêt significative, même si le modèle est spécifié incorrectement (c'est-à-dire que η_i n'est pas exactement linéaire en ce qui concerne les covariables), puisqu'il s'agit de l'approximation linéaire de l'approximation d'ordre un (en supposant que la fraction d'échantillonnage est négligeable) à $E(\mathbf{B} | y, X)$ est donnée par \mathbf{B} ou

$$\sum_{h=1}^H W_h \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h) \mathbf{B} = 0 \quad (7)$$

où $W_h = N_h/n_h$, $y_{hi} = \mathbf{B}^T x_{hi}$ et $\bar{y}_h = E(\mathbf{B}^T | y_h, X_h)$, ou \mathbf{B}^T peut être remplacé par $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}$, et les N_h peuvent être traités comme une loi multinomiale de taille N paramétrisée par les probabilités inconnues de la strate d'inclusion q_1, \dots, q_H avec, par exemple, une loi a priori de Dirichlet. Donc, dans l'exemple d'une régression linéaire, où $V(\mu_i) = \sigma^2$ et $g(\mu_i) = 1$, (7) se résout en

$$\mathbf{B} = E(\mathbf{B} | y, X) = \left[\sum_{h=1}^H W_h \sum_{i=1}^{n_h} x_{hi} x_{hi}^T \right]^{-1} \left[\sum_{h=1}^H W_h \sum_{i=1}^{n_h} x_{hi} x_{hi}^T \mathbf{B} \right] \quad (8)$$

Dans l'exemple de la régression logistique, où $V(\mu_i) = \mu_i(1 - \mu_i)$ et $g(\mu_i) = \mu_i^{-1}(1 - \mu_i)^{-1}$, $E(\mathbf{B} | y, X)$ s'obtient en résolvant, pour trouver les paramètres de régression de population B_j , $j = 1, \dots, p$,

$$\sum_{h=1}^H W_h \sum_{i=1}^{n_h} \frac{\exp(x_{hi} \mathbf{B})}{\exp(x_{hi} \mathbf{B}) + 1} = \sum_{h=1}^H W_h \sum_{i=1}^{n_h} \frac{\exp(x_{hi} \mathbf{B})}{\exp(x_{hi} \mathbf{B}) + 1} \quad (9)$$

On peut, pour cela, appliquer de simples méthodes numériques de calcul de racine comme la méthode de Newton. Dans le présent article, nous considérons quatre formes

1. Pente aléatoire échangeable (PAE) :

$$\mathbf{B}^h = (\beta_0^h, \dots, \beta^p)^T \text{ pour tous } h, G = I^H \otimes \sum. \quad (10)$$

2. Pente aléatoire autorégressive (PAA) :

$$\mathbf{B}^h = (\beta_0^h, \dots, \beta^p)^T \text{ pour tous } h, G = A \otimes \sum, A_{jk} = \rho^{j-k}, j, k = 1, \dots, H.$$

cela peut s'accomplir en créant un indice $h = 1, \dots, H$ de la probabilité d'inclusion (Little 1983, 1991); il pourrait s'agir d'une application bijection des statistiques d'ordre des poids des cas sur leurs classements, ou d'un « regroupement » préliminaire des poids des cas en utilisant, par exemple, les centiles 100/H des poids des cas. Les données sont alors modélisées par

$$y_{hi} | \theta_h \sim f(y_{hi}; \theta_h), i = 1, \dots, N_h$$

pour tous les éléments figurant dans la h^e strate d'inclusion, où θ_h tient compte d'une interaction entre le ou les para-

mètres du modèle θ et la strate d'inclusion h . Appliquer une loi a priori non informative à θ_h reproduit alors une analyse entièrement pondérée en ce qui concerne l'espérance de la loi prédictive a posteriori de $\tilde{Q}(Y)$.

Pour concrétiser ce qui précède, supposons que nous cherchions à estimer une moyenne de population $\tilde{Q}(Y) = \bar{Y} = N^{-1} \sum_{i=1}^N y_i$ d'après un échantillon à probabilités d'in-

clusion inégales avec un échantillonnage aléatoire simple dans les strates d'inclusion. En réécrivant l'équation sous la forme $\tilde{Q}(Y) = \sum_h h^e \bar{Y}_h$ où $\bar{Y}_h = N_h^{-1} \sum_{i=1}^{N_h} y_{hi}$ est la moyenne de la population de la strate d'inclusion et $P_h = N_h/N$, nous

avons

$$E(\bar{Y} | Y^{obs}) = \sum_h P_h E(\bar{Y}_h | Y^{obs}) =$$

$$N^{-1} \sum_h \{ n_h \bar{Y}_h | Y^{obs} + (N_h - n_h) E(\bar{Y}_h | Y^{obs}) \}$$

où \bar{Y}_h est décomposé en la moyenne de strate d'inclusion observée $\bar{Y}_h^{obs} = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ et la moyenne de strate d'inclusion non observée $\bar{Y}_h^{nob} = (N_h - n_h)^{-1} \sum_{i=1}^{N_h-n_h} (1 - I_{hi}) y_{hi}$. Si nous supposons que

$$y_{hi} | \mu_h, \sigma_h^2 \sim N(\mu_h, \sigma_h^2)$$

$$p(\mu_h, \sigma_h^2) \propto 1$$

alors

$$E(\bar{Y}_h^{nob} | Y^{obs}) =$$

$$E(E(\bar{Y}_h^{nob} | Y^{obs}) | Y^{obs}, \mu_h, \sigma_h^2) = E(\mu_h | Y^{obs}) = \bar{\mu}_h^{obs}.$$

et la moyenne prédictive a posteriori de la moyenne de population est donnée par la moyenne d'échantillon pondérée :

$$E(\bar{Y} | Y^{obs}) = \sum_h P_h E(\bar{Y}_h | Y^{obs}) =$$

$$N^{-1} \sum_{h=1}^H \sum_{i=1}^{N_h} I_{hi} y_{hi}^{obs} = N^{-1} \sum_{h=1}^H N_h \bar{Y}_h^{obs}$$

où $w_{hi} \equiv w_h = N_h/n_h$ pour tous les éléments observés dans la strate d'inclusion h . En outre, la moyenne pondérée sera l'espérance prédictive a posteriori de la moyenne de population pour toute loi hypothétique de X à condition que $E(Y_{hi} | \mu_h) = \mu_h$. Par contre, un simple modèle d'échan-

geabilité pour les données

donne $E(\bar{Y} | Y^{obs}) = n^{-1} \sum_{i=1}^n I_i y_i$. L'estimateur non pondéré de la moyenne, qui peut être gravement biaisé si l'échangeabilité n'est pas vérifiée, comme cela serait le cas s'il existait une association entre la probabilité d'inclusion et X .

2.2 Modèles de lissage des poids

Sous sa forme générale, la « méthode de lissage des poids » que nous proposons consiste à stratifier les données en fonction de la probabilité d'inclusion, puis d'utiliser un modèle hiérarchique pour effectuer la réduction par la voie d'un rééchantillonnage (shrinkage). Une description d'un tel modèle est donnée par

$$y_{hi} | \theta_h \sim f(y_{hi}; \theta_h) \quad (3)$$

$$\theta_h | M_h, \mu, R \sim N(\hat{y}_h(R), \hat{y}_h = g(M_h, \mu))$$

$$\mu, R | M_h \sim \Pi.$$

où $h = 1, \dots, H$ indice les probabilités d'inclusion de la plus élevée à la plus faible, $g(M_h, \mu)$ est une fonction liant l'information M_h provenant de la strate de probabilités d'inclusion et un paramètre de lissage μ au paramètre de distribution des données θ_h indicé par la strate d'inclusion, et Π est une loi uniforme ou faiblement informative d'un hyperparamètre (Little 2004). Les détails de spécification de la fonction de la vraisemblance et de la distribution a priori dépendent du paramètre de population étudié, du plan d'échantillonnage, des hypothèses concernant la loi de y_i et des compromis entre l'efficacité et la robustesse. Postuler un modèle d'échangeabilité sur les moyens de strates d'inclusion provenant de l'exemple précédent donne (Lazzeroni et Little 1998, Elliott et Little 2000)

$$y_{hi} | \theta_h \sim N(\theta_h, \sigma^2)$$

$$\theta_h | \mu, \tau^2 \sim N(\mu, \tau^2).$$

Si nous supposons pour le moment que σ^2 et τ^2 sont connus, nous avons

$$E(\bar{Y} | Y^{obs}) =$$

$$N^{-1} \sum_{h=1}^H \{ n_h \bar{y}_h^{obs} + (N_h - n_h) E(\mu_h | Y^{obs}) \}$$

où $E(\mu_h | Y^{obs}) = w_h \bar{y}_h + (1 - w_h) \tau^2$ pour $w_h = \tau^2 n_h / (\tau^2 n_h + \sigma^2)$ et $\bar{y}_h = (\sum_{i=1}^{n_h} y_{hi} / \tau^2 + \sigma^2)^{-1} \sum_{i=1}^{n_h} y_{hi} / (n_h / \tau^2 + \sigma^2)$. À mesure que $\tau^2 \rightarrow \infty$, $w_h \rightarrow 1$ de sorte que $E(\bar{Y} | Y^{obs}) = \sum_h P_h \bar{y}_h$, donc, une loi a priori uniforme

uniquement de X^{obs} , alors le mécanisme d'échantillonnage est dit « ignorable » (Rubin 1987), ce qui équivaut à la terminologie standard des données manquantes (les éléments non observés de la population peuvent être considérés comme manquants par conception). Sous des plans d'échantillonnage ignora- bles, $d(\theta|\phi) = d(\theta)$ et $p(I|X, \theta, \phi) = p(I|X^{obs}, \phi)$, et donc (1) se réduit à

$$\int \frac{p(X^{nob} | X^{obs}, \theta) d(\theta) d(X^{obs} | \theta) d\theta dX^{nob}}{p(X^{nob} | X^{obs}, \theta) d(X^{obs} | \theta) d\theta} = p(X^{nob} | X^{obs}), \quad (2)$$

ce qui permet de faire l'inférence au sujet de $\mathcal{Q}(X)$ sans modéliser explicitement le paramètre I d'inclusion dans l'échantillonnage (Ericson 1969, Holt et Smith 1979, Little 1993, Rubin 1987, Skinner, Holt et Smith 1989). Les plans d'échantillonnage non informatifs représentent un cas particulier des plans d'échantillonnage ignorables équivariant aux mécanismes de création de données manquant entière- ment au hasard qui sont un cas particulier des mécanismes de création de données manquant au hasard.

Dans les conditions de régression, où on souhaite faire des inférences au sujet des paramètres qui régissent la loi de Y sachant les covariables fixes et connues X , (1) devient

$$p(X^{nob} | X^{obs}, X, I) = \frac{\int \int p(X^{nob} | X^{obs}, X, \theta, \phi) \times \frac{p(I | X, X^{nob}, X^{obs}, \theta, \phi) d\theta d\phi dX^{nob}}{p(I | X, X, \theta, \phi) d(\theta) d(X^{obs} | X, \theta, \phi) d\theta d\phi}}{p(X^{nob} | X^{obs}, X, I)}$$

qui se réduit à

$$p(X^{nob} | X^{obs}, X) = \frac{\int p(X^{nob} | X^{obs}, X, \theta, \phi) d(X^{obs} | X, \theta, \phi) d\theta d\phi}{\int p(X^{nob} | X^{obs}, X, \theta, \phi) d(X^{obs} | X, \theta, \phi) d\theta d\phi}$$

si, et uniquement si, I dépend seulement de (X^{obs}, X) , dont la dépendance à l'égard de X uniquement est un cas particulier. Donc, si l'on souhaite faire une inférence au sujet d'un paramètre de régression $\mathcal{Q}(X, X)$, alors un plan d'échantillonnage non informatif ou, plus généralement, ignorable peut permettre que les probabilités d'inclusion soient une fonction des covariables fixes.

2.1 Adaptation à des probabilités d'inclusion inégales

Maintenir l'hypothèse d'ignorabilité du mécanisme d'échantillonnage oblige souvent à tenir compte du plan d'échantillonnage dans la structure du modèle de vraisemblance et du modèle a priori. Dans le cas des plans d'échantillonnage avec probabilités d'inclusion inégales,

Représentons les données de population pour une population comptant $i = 1, \dots, N$ unités par $X = (X_1, \dots, X_N)$, et la variable indicatrice d'échantillonnage $I = (I_1, \dots, I_N)$ où $I_i = 1$ si le i^{e} élément est échantillonné et 0 autrement. Comme l'inférence fondée sur le plan de sondage, l'inférence bayésienne se concentre sur les quantités étudiées de population $\mathcal{Q}(X)$, comme les moyennes de la population $\bar{Q}(X)$ ou les paramètres de régression par les moindres carrés de population $\bar{Q}(X, X) = \min_{B_0, B_1} \sum_{i=1}^N (y_i - B_0 - B_1 x_i)^2$. Contrairement à l'inférence fondée sur le plan, mais en accordance avec la plupart des autres domaines de la statistique, on postule pour les données de population X un modèle ayant la forme d'une fonction des paramètres $\theta : X \sim f(X|\theta)$. L'inférence au sujet de $\mathcal{Q}(X)$ est faite en se fondant sur la loi prédictive a posteriori de $p(X^{nob} | X^{obs}, I)$, où X^{nob} consiste en les éléments de X_i pour lesquels $I_i = 0$:

Si nous supposons que ϕ et θ sont indépendants a priori et si la loi de l'indicatrice d'échantillonnage I est indépendante de X , le plan d'échantillonnage est dit « non confon- du » ou « non informatif »; si la loi de I dépend

$$p(X^{nob} | X^{obs}, I) = \frac{\int \int \int p(X^{nob} | X^{obs}, \theta, \phi) p(I | X, \theta, \phi) d\theta d\phi dX^{nob}}{\int \int p(X^{nob} | X^{obs}, \theta, \phi) p(I | X, \theta, \phi) d\theta d\phi} \quad (1)$$

où $p(I | X, \theta, \phi)$ modélise l'indicatrice d'inclusion.

consiste à déterminer l'EQM empirique au niveau de réduction w_i où le poids réduit est $w_i' = w_i I(w_i' > w_i) + w_i / I(w_i' > w_i)$, $i = 1, \dots, n$ sous l'hypothèse que l'estimation entièrement pondérée est sans biais pour la moyenne réelle. En pratique, on envisage une série de niveaux de réduction $t = 1, \dots, T$, où $t = 1$ correspond aux données non pondérées ($w_i = \min(w_i')$) et $t = T$, aux données entièrement pondérées ($w_i = \max(w_i')$), et θ_t est la valeur de la statistique en utilisant les poids réduits au niveau t . Le niveau de réduction choisi est alors donné par $w_0 = w_{t^*}$, où $t^* = \text{argmin}_t (EQM_t)$ pour $EQM_t' = (\theta_t' - \theta_t)^2 + V(\theta_t)$.

La littérature sur le calage décrit des techniques qui ont été mises au point en vue de permettre que des ajustements (raking) généralisés soient bons pour éviter la construction de poids extrêmes (Deville et Särndal 1992, Folsom et Singh 2000). Beaumont et Alavi (2004) étendent cette notion à l'élaboration d'estimateurs axés sur la réduction des poids de valeur élevée d'observations fortement influentes ou aberrantes. Bien que l'imposition de ces bornes réduit les poids extrêmes pour les ramener à une valeur seuil fixe, le choix de ce seuil demeure arbitraire.

Une autre approche des méthodes de réduction direct des poids a été élaborée dans la littérature sur l'inférence bayésienne en population fine (Elliott et Little 2000, Holt et Smith 1979, Ghosh et Meeden 1986, Little 1991, 1993, Lazzeroni et Little 1998, Rizzo 1992). Ces approches tiennent compte des probabilités d'inclusion inégales en considérant les poids des cas comme des variables de stratification à l'intérieur des strates définies par la probabilité d'inclusion. Ces « strates d'inclusion » peuvent correspondre aux strates formelles définies par un plan de sondage disproportionnel stratifié, ou peuvent être des « pseudo-strates » fondées sur des poids regroupés dérivés de la sélection, de la poststratification et (ou) des redressements pour la non-réponse. Des estimations pondérées standard sont alors obtenues quand les moyennes de strate de poids des résultats de l'enquête sont traitées comme des effets fixes et que la réduction des poids est réalisée en considérant les moyennes de strate de poids sous-jacentes comme des effets aléatoires. Ces méthodes tiennent compte de la présence éventuelle de données « partiellement pondérées » qui utilisent les données proprement dites pour moduler comme il convient le compromis entre le biais et la variance, et permettent aussi que l'estimation et l'inférence à partir de données recueillies sous des plans de sondage avec probabilités d'inclusion inégales soit fondées sur des modèles utilisés dans d'autres domaines d'estimation et d'inférence statistiques.

Le présent article étend ces modèles à effets aléatoires, que nous appelons modèles « de lissage des poids », afin

score pour les paramètres de régression sous le modèle de régression de superpopulation hypothétique si nous disposons de données observées pour l'ensemble de la population. La convergence par rapport au plan impliqué que la différence entre la quantité cible de population et l'estimation calculée d'après l'échantillon tend vers zéro quand la taille de l'échantillon et la taille de la population augmentent concomitamment, ou que les différences tendent, en moyenne, vers zéro par échantillonnage répété de la population, où les échantillons sont sélectionnés de manière identique à partir de $t \rightarrow \infty$ répétitions de la population : voir Särndal (1980) ou Isaki et Fuller (1982). Si les observations sont en grappes, plus de soins doivent être mis à élaborer des estimateurs EPMV convergents sous le plan, quoique les plans hiérarchiques à plusieurs degrés permettent d'approcher les estimations de la log-vraisemblance dans des conditions de recensement à l'aide d'équations de score pondérées si l'on veille à tenir compte du fait que les tailles d'échantillons intra-grappe sont habituellement faibles et le demeurent même si le nombre de grappes augmente (Pfefferman, Skinner, Holmes, Goldstein et Rabash 1998, Korn et Graubard 2003).

Bien que les EPMV soient populaires, à cause de leur convergence sous le plan, cette propriété est obtenue au prix d'un accroissement de la variance. Cet accroissement peut englober la réduction du biais, de sorte qu'effectivement, l'EQM augmente dans une analyse pondérée. Cela risque surtout de se produire si a) la taille d'échantillon est faible, b) les écarts entre les probabilités d'inclusion sont grands ou c) le modèle est approximativement correctement spécifié et l'échantillonnage est approximativement non informatif. L'approche qui est peut-être la plus courante pour traiter ce problème est la réduction des poids (Porter 1990, Kish 1992, Alexander, Dahl et Weidman 1997), qui consiste à fixer la valeur des poids plus grands qu'une certaine valeur w_0 à cette valeur w_0 . Habituellement, w_0 est choisie de manière ponctuelle - disons égale à trois ou à six fois le poids moyen - sans se soucier de savoir si le seuil de réduction choisi est optimal en ce qui concerne l'EQM. Donc, le biais est introduit pour réduire la variance, avec l'objectif d'une réduction globale de l'EQM.

D'autres méthodes fondées sur le plan de sondage ont été envisagées dans la littérature. Porter (1990) discute de méthodes systématiques en vue de choisir w_0 , y compris des méthodes de distribution des poids consiste à supposer que les poids suivent une loi bêta inverse et à l'aide d'estimateurs par la méthode des moments, et les poids provenant de la queue supérieure de la distribution, disons où $1 - F(w_0) < 0,01$, sont réduits à la valeur w_0 telle que $1 - F(w_0) = 0,01$. La méthode de réduction de l'EQM

Réduction bayésienne des poids pour les modèles de régression linéaire généralisée

Michael R. Elliott

Résumé

Dans les sondages où les unités ont des probabilités inégales d'inclusion dans l'échantillon, les associations entre la probabilité d'inclusion et la statistique d'intérêt peuvent causer un biais. Des poids égaux à l'inverse de la probabilité moyenne de valeur élevée qui peuvent introduire une variabilité indésirable dans les statistiques telles que l'estimateur de la somme de valeur élevée à la somme des poids non réduits, ce qui réduit la variabilité au prix de l'introduction d'un certain biais. La plupart des approches ordinaires sont ponctuelles en ce sens qu'elles n'utilisent pas les données en vue d'optimiser le compromis entre le biais et la variance. Les approches dites par les données qui sont décrites dans la littérature sont un peu plus efficaces que les estimateurs entièrement pondérés. Dans le présent article, nous élaborons des méthodes bayésiennes de réduction des poids d'estimateurs par la régression linéaire et par la régression linéaire généralisée sous des plans de sondage avec probabilités d'inclusion inégales. Nous décrivons une application à l'estimation du risque de blessure chez les enfants installés sur le siège arrière dans les camionnettes compactes à l'aide des données de la Partners for Child Passenger Safety surveillance survey.

Mots clés : Sondage; poids de sondage; winsorisation des poids; inférence bayésienne à la population; lissage des poids; modèles linéaires généralisés mixtes.

1. Introduction

Lors de l'analyse de données provenant d'échantillons sélectionnés avec probabilités différentes, on utilise souvent comme poids des cas les inverses des probabilités d'inclusion afin de réduire ou d'éliminer le biais dans les estimateurs des quantités de population d'intérêt. Le remplacement des moyennes et des totaux implicites dans les statistiques par leurs équivalents pondérés par les poids des cas produit des estimateurs linéaires sans biais et des estimateurs non linéaires asymptotiquement sans biais des valeurs de population (Binder 1983). Les poids des cas peuvent aussi intégrer des ajustements pour la non-réponse, qui habituellement sont égaux à l'inverse de la probabilité estimée de réponse (Gelman et Carlin 2002, Oh et Scheuren 1983), ou des ajustements par calage, qui contraignent les poids des cas à être égaux à des totaux connus de pondération, soit conjointement, comme dans la poststratification ou l'estimation par la régression généralisée, soit aux marges, comme dans l'estimation par calage sur marges (raking) généralisée (Deville et Särndal 1992, Isaki et Fuller 1982).

L'utilisation des poids de sondage pour la production de statistiques descriptives, comme les moyennes et les totaux, d'après des plans avec probabilités d'inclusion inégales ne suscite guère de débat. Cependant, lorsqu'il s'agit d'estimer des quantités « analytiques » (Cochran 1977, page 4) axées

sur les associations entre, par exemple, les facteurs de risque et les résultats en matière de santé au moyen de modèles linéaires ou linéaires généralisés, la décision d'utiliser les poids de sondage est moins catégorique (voir Korn et Graubard 1999, pages 180-182). Dans des conditions de régression, des différences entre les estimateurs pondérés et non pondérés de la pente de la droite de régression peuvent survenir parce que le modèle de données est spécifié incorrectement ou qu'il existe une association entre les erreurs résiduelles et (ou) la probabilité d'inclusion (l'échantillon est informatif). Si le modèle de données est spécifié incorrectement, une option consiste à améliorer la spécification du modèle. Cependant, il est parfois difficile de déterminer la forme fonctionnelle exacte; ou bien, il se peut que l'erreur de spécification soit très modeste, mais que le plan de sondage l'amplifie, ou bien, une approximation du modèle réel pourrait être souhaitée pour simplifier l'explication (approximation linéaire d'une tendance quadratique). Dans le cas de l'échantillonnage informatif ou non ignorable, les poids de sondage peuvent être nécessaires pour obtenir des estimateurs convergents des paramètres de régression (Korn et Graubard 1995). De manière plus formelle, les estimateurs entièrement pondérés des paramètres de régression sont des estimateurs du « pseudo-maximum de vraisemblance » (EPMV) (Binder 1983, Pfeffermann 1993) en ce sens qu'ils sont « convergents sous le plan » pour les EMV qui résoudre les équations de

- Nous avons aussi comparé visuellement, dans le cas de notre exemple, les estimations des erreurs-types des coefficients obtenues sous le plan de sondage (en tenant compte de la mise en grappes au niveau de l'UPÉ et à des niveaux inférieurs), ainsi que sous une modification de l'approche robuste fondée sur un modèle (tenant compte de la mise en grappes au niveau de l'individu ou à un niveau inférieur) pour les modèles 1 et 2. Nous n'avons observé que des différences faibles, qui indiquaient l'absence d'effets de grappe à un niveau d'aggrégation supérieur au niveau de l'individu pour les données en question. Nous avons également calculé les estimations des erreurs-types en supposant que les périodes chez une même personne étaient indépendantes et de nouveau constaté des différences faibles seulement par rapport à celles obtenues suivant l'approche Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.
- Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, 47, 939-956.
- Lavigne, M., et Michaud, S. (1998). General Aspects of the Survey of Labour and Income Dynamics. Document de travail, Statistique Canada, 75F0002M No. 98-05.
- Lin, D.Y. (1994). Cox regression analysis of multivariate failure time data: a marginal approach. *Statistics in Medicine*, 13, 2233-2247.
- Lin, D.Y. (2000). On Fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.
- Lin, D.Y., et Wei, L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.
- Prentice, R.L., Williams, B.J. et Peterson, A.V. (1981). On the regression analysis of multivariate failure data. *Biometrika*, 68, 373-379.
- Roberts, G., et Kovacević, M. (2001). New research problems in surveys. *Proceedings of the Survey Methods Section of the Statistical Society of Canada*, 111-116.
- Wei, L.J., Lin, D.Y. et Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.
- Williams, R.L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.
- Clayton, D., et Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (avec discussion). *Journal of the Royal Statistical Society, Séries A*, 1985, 148, 82-117.
- Biosseld, H.-P., et Hamerle, A. (1989). Using Cox models to study multiepisodic processes. *Sociological Methods and Research*, 17, 4, 432-448.
- Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79, 139-147.
- Binder, D.A. (1983). On the variance of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-291.

Bibliographie

Remerciements

Nous remercions Normand Lanlet et Xuelin Zhang de leurs commentaires constructifs concernant une version antérieure du manuscrit. Nous remercions également le rédacteur adjoint et les examinateurs de leurs commentaires et suggestions qui ont accru considérablement la lisibilité du manuscrit.

Nous avons aussi comparé visuellement, dans le cas de notre exemple, les estimations des erreurs-types des coefficients obtenues sous le plan de sondage (en tenant compte de la mise en grappes au niveau de l'UPÉ et à des niveaux inférieurs), ainsi que sous une modification de l'approche robuste fondée sur un modèle (tenant compte de la mise en grappes au niveau de l'individu ou à un niveau inférieur) pour les modèles 1 et 2. Nous n'avons observé que des différences faibles, qui indiquaient l'absence d'effets de grappe à un niveau d'aggrégation supérieur au niveau de l'individu pour les données en question. Nous avons également calculé les estimations des erreurs-types en supposant que les périodes chez une même personne étaient indépendantes et de nouveau constaté des différences faibles seulement par rapport à celles obtenues suivant l'approche Klein, J.P. (1992). Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48, 795-806.

postulée explicitement et prise en compte dans la formule d'estimation de la variance et où, dans l'approche fondée sur le plan de sondage, les périodes chez un même individu sont traitées comme une grappe dans l'estimation de la variance sous le plan de sondage.

Nous avons appliqué l'approche fondée sur le plan de sondage à trois modèles de type à risques proportionnels. L'un tenait compte de risques de base non spécifiés différents et de coefficients différents pour chaque ordre de période. Le deuxième tenait encore compte de risques de base non spécifiés différents pour divers ordres de période, mais exigeait que les coefficients soient les mêmes pour tous les ordres. Le troisième était un simple modèle à période unique. Nous avons constaté que la façon dont l'information sur l'ordre de la période était utilisée avait une incidence sur les résultats de l'ajustement de nos modèles. Une comparaison visuelle des estimations des coefficients et des estimations des risques de base cumulatifs pour les modèles 1 et 2 indiquait des résultats différents. Comme l'a suggéré l'un des examinateurs, il serait bon d'élaborer un test formel en vue de déterminer si les coefficients diffèrent effectivement en fonction de l'ordre de la période (comme dans le modèle 1), étant donné des risques de base pouvant varier selon l'ordre de la période. Il est, en fait, facile de produire un tel test, de la façon suivante. Soit $\gamma = (B_1, B_2, \dots, B_K)'$ le vecteur des K vecteurs de coefficients du modèle 1, où chacun est de longueur p , et soit $z_j(t) = (0', 0', \dots, 0', x_j(t), 0', \dots, 0')'$ le vecteur de longueur pK pour la j^{e} période du j^{e} individu où la j^{e} composante de ce vecteur contient le vecteur des covariables $x_j(t)$. Alors, le

modèle 1 peut être exprimé par

$$h_j^*(t|z_j(t)) = \lambda_{j0}(t) e^{z_j(t)'\gamma},$$

qui a la forme générale des risques de base variant avec l'ordre de la période, mais ayant un vecteur de coefficients fixe. Un test de la convergence des coefficients se rapportant à chaque ordre de période, c'est-à-dire $H_0: B_1 = B_2 = \dots B_K$, équivaut à tester $H_0: C\gamma = 0$ où C est la matrice $C = I_p \otimes [I_{K-1} - I_K]$ de dimensions $(K-1)p \times Kp$. Étant donné une estimation $\hat{\gamma}$ de γ et une estimation $\hat{V}(\hat{\gamma})$ de la matrice de covariance de $\hat{\gamma}$, obtenue directement à la section 4 pour le modèle 2, une statistique de Wald peut être calculée pour tester l'hypothèse. Si l'hypothèse n'est pas rejetée, on peut conclure qu'un modèle à coefficients constants sur l'ordre de la période (mais à risques de base variables en fonction de l'ordre de la période) semble être aussi bien ajusté aux données qu'un modèle où les risques de base et les coefficients varient en fonction de l'ordre de la période. D'autres mesures de l'adéquation du modèle devraient également être faciles à élaborer dans le cadre d'estimations sous le plan de sondage.

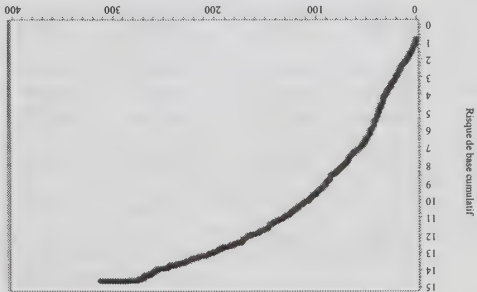


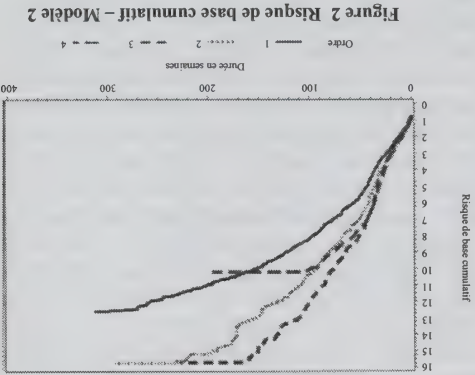
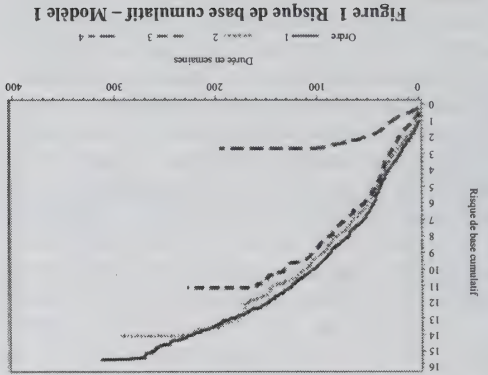
Figure 3 Risque de base cumulatif – Modèle 3

6. Conclusion

Nous avons étudié le problème de l'analyse de périodes multiples en considérant deux approches générales pour traiter le manque d'indépendance entre les temps de sortie, à savoir une approche robuste basée sur un modèle et une approche basée sur le plan de sondage. La première consiste à estimer les paramètres du modèle en supposant que les périodes sont indépendantes, puis à corriger la matrice de covariance naïve de façon à tenir compte des dépendances intra-individu postulées par le chercheur. Cette approche ne tient pas compte de la mise en grappes éventuelle entre individus (ou, en fait, de toute mise en grappes qui pourrait avoir lieu à un niveau d'agrégation plus élevé que l'individu) ni des probabilités inégales de sélection des individus (par exemple, nous avons montré comment la méthode pourrait être étendue afin d'inclure les poids de sondage). La deuxième approche définit les coefficients du modèle comme des paramètres de population finie. Ces paramètres sont ensuite estimés en tenant compte des probabilités de sélection éventuellement inégales des individus. Une méthode d'estimation de la variance sous le plan de sondage qui tient compte des corrélations éventuelles entre individus dans la même UPE rend compte automatiquement des dépendances non spécifiées des périodes à des niveaux inférieurs à l'UPE, comme les dépendances intra-individu. Dans le cas des échantillons de grande taille, cette inférence sous le plan de sondage s'étend directement à la super-population à partir de laquelle la population finie a été hypothétiquement générée. Le défaut de la première approche est qu'elle ignore totalement la possibilité d'une mise en grappes entre individus. Un inconvénient éventuel de la deuxième approche, comme nous l'avons appliquée, est qu'elle repose sur l'hypothèse de l'échantillonnage aléatoire simple d'individus, où, dans le cas de l'approche robuste fondée sur un modèle, la dépendance entre les périodes chez un même individu est

Tableau 2 Coefficients β estimés pour les trois modèles

	Modèle 1	Modèle 2	Modèle 3
SEX (F)	Ordre 1	Ordre 2	Ordre 3
M	0,4417	0,3781	0,4435
EDUCLEV (É)	-0,4561	-0,5234	-0,4128
F	-0,2330	-0,2700	-0,2436
MF	-0,0744	-0,1060	-0,0684
M	-0,1442	-0,1290	-0,1357
MARST (M)	0,0985	-0,0894	0,0328
Celibataire	0,0704	0,2752	0,3413
TYPIBEND (Congédié(e))			0,1579
Départ volontaire			0,0490
OCCUPATION (Autre)	0,1592	-0,1364	0,0903
Professionnels	-0,0265	-0,2930	-0,0971
Administration	-0,0211	-0,2175	-0,0410
Secleur primaire	-0,0003	-0,0994	-0,0093
Fabrication	0,1290	-0,1862	0,0490
Construction	-0,0003	-0,0994	-0,0088
FIRMSIZE (1 000+)	-0,0027	-0,0097	0,0441
<20	0,0388	0,0881	0,0928
20 à 99	0,0436	-0,0905	0,0214
100 à 499	-0,0006	0,0153	-0,0005
500 à 999	-0,2903	-0,5414	-0,3693
ATSCN (Non)			-0,3743
Oui	-1,0832	-1,1516	-1,1205
Revenu familial par			-1,1266
personne (10 000 \$ et			
moins)			
10 000 \$ à 20 000 \$	0,1294	0,1802	0,1345
20 000 \$ à 30 000 \$	0,1644	0,3611	0,2241
30 000 \$ et plus	0,1712	0,3916	0,2280
AGE	-0,0491	-0,0311	-0,0424
Périodes dans l'ensemble			
de risques	8 386	4 255	1 300
Censurées	1 913	759	281
Achevées	6 473	3 496	1 019
Les valeurs significatives au seuil de signification de 5 % sont en caractères gras.			



L'examen visuel des fonctions de survie de Kaplan-Meier (non présentes) pour les périodes de chaque ordre révélait que, à mesure que croissait l'ordre, la valeur de la fonction de survie à n'importe quel temps fixe t diminuait, indiquant que les premières périodes sont les plus longues parmi les périodes achevées et que la durée d'une période multiple est d'autant plus courte que son ordre est élevé. Cette constatation est vraisemblablement une conséquence de la durée de vie limitée du panel, en ce sens qu'un individu comptant un plus grand nombre de périodes durant l'intervalle de temps de six ans donné est susceptible de présenter des périodes plus courtes.

5.4 Ajustement des modèles selon une approche fondée sur le plan

Comme nous l'avons mentionné plus haut, notre exemple est simplement une illustration de l'approche fondée sur le plan de sondage d'ajustement des modèles à risques proportionnels à des données sur des événements multiples provenant d'une enquête à plan de sondage complexe. Donc, nous consacrons peu de temps ici à discuter de la façon d'évaluer l'adéquation de ces modèles, dont l'adéquation des hypothèses de proportionnalité dans chacun des modèles ou le fait de savoir si un type de modèle est aussi bien ajusté qu'un autre.

Les coefficients estimés d'après l'ajustement des trois modèles aux données de l'EDTR sont présentées au tableau 2. Les valeurs des coefficients significatives au seuil de 5 % sur la base de tests t individuels sont en caractères gras.

Le modèle 1 est conditionnel à l'ordre de la période et comprend l'ajustement de quatre modèles distincts aux données provenant des quatre ordres de période différents. Comme le montre le tableau 2, les coefficients des variables SEXE, AGE et au moins une catégorie de la variable de revenu familial sont significatifs pour les périodes de tous ordres, quoique leur grandeur estimée varie en fonction de l'ordre de la période. Les coefficients estimés pour AGE sont négatifs, mais diminuent de grandeur à mesure que l'ordre de la période augmente, tandis qu'aucune tendance n'est discernable pour les coefficients estimés pour les deux autres variables. Les variables EDUCLEV, PARTJB et ATSCH ont des coefficients significatifs pour les périodes d'ordre 1, 2 et 3, mais non pour les périodes d'ordre 4. Ce résultat peut être attribué, en partie, à la petite taille d'échantillon pour les quatrièmes périodes. Pour chacune des trois autres variables du modèle (MARST, OCCUPATION et FIRMSIZE), un coefficient n'était significatif que pour un seul ordre de période.

Pour le modèle 2, les coefficients sont contraints d'être les mêmes pour tous les ordres de période. Comme le

montre le tableau 2, numériquement, un grand nombre de valeurs des coefficients estimés, mais pas toutes, sont comprises entre les estimations calculées pour la première et pour la deuxième période à l'aide du modèle 1, ce qui pourrait être dû au fait qu'une proportion élevée de l'échantillon correspond aux événements de ces ordres. Toutes les variables, sauf OCCUPATION, ont un coefficient significatif. Les erreurs-types des coefficients sont plus faibles dans le cas du modèle 2 que dans celui du modèle 1.

Le modèle 3 est un modèle à période unique avec un seul ensemble de coefficients et une seule fonction de risque de base. Les coefficients estimés du modèle sont semblables aux estimations obtenues à l'aide du modèle 2.

Les fonctions de risque de base cumulatif estimées pour les modèles 1 à 3 sont illustrées aux figures 1 à 3 respectivement. Dans tous les cas, pour les durées allant jusqu'à environ 50 semaines, les fonctions ont une forme concave, ce qui sous-entend qu'il existe une dépendance positive par rapport au temps du taux de sortie (autrement dit, la probabilité de sortie est d'autant plus élevée que la période est longue). Pour les durées de plus de 50 semaines, la forme devient convexe, ce qui suggère une dépendance négative par rapport au temps dans le cas des périodes plus longues. À la figure 1, la position de la fonction de risque de base cumulatif estimée varie selon l'ordre de la période, la courbe pour les périodes d'ordre 1 étant la plus élevée et celle pour les périodes d'ordre 4, la plus basse. À la figure 2, pour le modèle 2, les positions des diverses courbes ne suivent pas l'ordre des périodes. Cette différence observée entre les figures 1 et 2 pourrait servir de diagnostic visuel indiquant qu'une étude plus approfondie est nécessaire afin d'évaluer lequel des modèles 1 ou 2 est un meilleur descripteur des données, puisque les coefficients estimés ont une incidence sur les risques de base estimés.

5.5 Comparaison aux estimations robustes modifiées de la variance sous le modèle

Comme il est décrit à la section 5.2, les estimations robustes modifiées de la variance sous le modèle tiennent compte de la corrélation éventuelle entre les périodes chez un même individu, sous l'hypothèse d'indépendance entre les individus. La comparaison, pour les modèles 1 et 2, des estimations des erreurs-types obtenues par cette approche aux estimations des erreurs-types sous le plan de sondage n'a révélé que des écarts faibles. Il semble donc que les estimations sous le plan reflètent toute corrélation entre les périodes chez un même individu et qu'il n'y ait pas de dépendance supplémentaire au-delà du niveau individuel dans notre exemple.

périodes admissibles est représentée par un ensemble de lignes dont chacune correspond à une période. Bien qu'une ligne contienne le temps d'entrée dans la période t_1 et le temps de sortie de la période t_2 ou le temps de censure t_c , pour l'analyse, la durée est toujours considérée sous la forme $(0, t_2 - t_1)$ ou $(0, t_c - t_1)$. Les covariables examinées sont reliées à chaque ligne. Sont également reliées à chaque ligne le poids longitudinal de 1998 et les identificateurs de la strate et de l'UPÉ de la personne dont la période est décrite par l'enregistrement en question.

5.2 Analyse

Pour les besoins de l'illustration, nous avons limité l'analyse aux quatre premières périodes, si bien que tous les individus échantillonnés présentant des périodes admissibles sont inclus dans l'analyse, mais que les enregistrements de période survenue après la quatrième ne sont pas pris en considération à cause de leur faible nombre dans l'échantillon.

Nous avons estimé les coefficients et leurs variances pour les trois modèles par les méthodes fondées sur le plan de sondage décrites à la section 4 en utilisant la procédures « SURVIVAL » du logiciel SUDAAN version 8. Pour chacun des trois modèles, nous avons spécifié un plan de sondage stratifié avec tirage des UPÉ avec remise (c'est-à-dire $DESIGN = WR$). Les trois modèles ont été ajustés au même nombre de périodes (16 307). Pour chaque modèle, nous avons ensuite calculé les fonctions de risque de base cumulatif empiriques en utilisant une approche produite limite (voir Kalbfleisch et Prentice (2002), pages 114-116) telle qu'elle est implémentée dans la procédure SURVIVAL de SUDAAN.

L'approche robuste sous le modèle pour les périodes multiples décrites à la section 3.1 comporte un ajustement des estimations de la variance en vue de tenir compte de l'interdépendance éventuelle des périodes chez un même individu, en supposant que les périodes provenant d'individus différents sont indépendantes; cependant, dans cette approche, aucune mesure n'est prise pour tenir compte des probabilités inégales de sélection des individus échantillonnés, que ce soit dans les estimations des coefficients ou les estimations de la variance. Afin de le faire, pour les modèles 1 et 2, nous avons également utilisé la procédure SURVIVAL de SUDAAN version 8 pour estimer les variances des estimations pondérées des coefficients, où nous avons émis l'hypothèse d'indépendance inter-individus dans les périodes, mais nous avons tenu compte de la corrélation intra-individu éventuelle des périodes. Pour cela, nous avons spécifié un plan d'échantillonnage non stratifié avec sélection des grappes avec remise en précisant que chaque individu formait sa propre grappe. Les hypothèses de dépendance sont les mêmes que celles utilisées par Lin (1994), mais nous avons tenu compte de l'utilisation de pondérations dans l'estimation des coefficients et des variances. Nous appellerons ces estimations des variances « estimations robustes modifiées de la variance sous le modèle des estimations pondérées des coefficients ».

5.3 Certaines statistiques descriptives

La durée moyenne estimée d'une période achevée est de 33,3 semaines, tandis que la durée moyenne estimée de la partie observée d'une période censurée (inachevée) est de 48,5 semaines.

Tableau 1 Dénombrement des individus du panel de six ans de l'EDTR ayant des périodes de chômage débutant entre janvier 1993 et décembre 1998, selon le nombre total de périodes et l'ordre de la période (A-achevée, I-inachevée)

Individus selon le nombre de périodes	Périodes par ordre					S ⁺ +				
	Première	Deuxième	Troisième	Quatrième	S ⁺	A	I	A	I	S ⁺
1 période	4 141	2 221	1 920	-	-	-	-	-	-	-
2 périodes	1 915	1 915	1 154	761	-	-	-	-	-	-
3 périodes	1 044	1 044	1 044	612	432	-	-	-	-	-
4 périodes	629	629	629	-	-	348	281	-	-	-
5 périodes et plus	672	672	-	672	-	672	-	1 158	415	415
Total	8 401	6 481	1 920	3 499	761	1 913	432	1 020	281	1 158
415										

données provenant des personnes qui ont répondu au dernier cycle du panel ont été incluses dans les analyses. Un bon résumé des questions relatives au plan d'échantillonnage de l'EDTR est donné dans Lavigne et Michaud (1998). Un examen des questions en rapport avec les études des périodes de chômage basées sur l'EDTR est donné dans Roberts et Kovačević (2001).

L'état d'intérêt est « être en chômage », défini ici comme l'état d'intérêt après le 1^{er} janvier 1993 ont été incluses, puisque le 31 janvier 1993 est la date de début des observations auprès du panel. Les périodes de chômage qui n'étaient pas achevées à la fin de la période d'observation (31 décembre 1998) ont été considérées comme étant censurées. Les dénombrements d'échantillon des individus concernés des périodes admises et des périodes en fonction de leur ordre sont présentés au tableau 1. Brevèvement, 17 880 périodes ont été dénombrées auprès de 8 401 membres du panel longitudinal. Environ la moitié des individus échantillonnés (4 260) devenus chômeurs durant cette période ont connu deux périodes de chômage ou plus. En tout, 3 809 périodes sont demeurées inachevées à cause de la cessation du panel.

Sur une longue liste de covariables disponibles, nous n'en avons choisies que dix. La variable de sexe [SEX] de l'individu longitudinal est la seule qui demeure constante au cours des diverses périodes. Quatre variables ont des valeurs enregistrées à la fin de l'année durant laquelle la période a commencé, à savoir le niveau de scolarité [EDUCLEV] avec quatre catégories (faible, moyen-faible, moyen, élevé), l'état matrimonial [MARST] avec trois catégories (célibataire, marié(e)/union de fait, autre), le revenu familial par personne (en dollars canadiens) avec quatre catégories (<10 000, de 10 000 à 20 000, de 20 000 à 30 000, 30 000 et plus) et l'âge [AGE] (en années). Trois variables ont les valeurs correspondant à l'emploi avec mise à pied qui a précédé la période, à savoir le type de fin d'emploi [TYPBEND] avec deux catégories (congé(é) et départ volontaire), la profession [OCCUPATION] avec six catégories (professionnel, administration, secteur primaire, fabrication, construction et autre) et la taille de l'entreprise [FIRMSIZE] avec cinq catégories (<20, de 20 à 99, de 100 à 499, de 500 à 999, et 1 000 et plus employés). Deux variables binaires représentent la situation durant la période, à savoir travailler à temps partiel [PARTTJB] et être aux études [ATSCHE].

L'ensemble de données a été préparé selon le mode de «dénombrement» où chaque individu représentant des

L'ensemble de données que nous utilisons pour l'illustration provient du premier panel de six ans (1993 à 1998) de l'enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada. Dans ce panel, environ 31 000 personnes sélectionnées dans environ 15 000 ménages ont été suivies pendant six ans par la voie d'interviews annuelles. Certaines personnes sont sorties de l'échantillon au cours des temps pour diverses raisons, tandis que quelques autres, après avoir manqué une ou plusieurs interviews, ont recommencé à participer au panel. Une pondération compliquée des personnes répondant à l'EDTR chaque année tient compte des divers types d'érosion du panel, de sorte que chaque répondant durant une année particulière soit pondéré en fonction de la population de référence pertinente de 1993. Cette approche produit une pondération longitudinale distincte pour chaque cycle (c'est-à-dire année) de collecte de données. Pour la présente analyse, nous avons utilisé les poids longitudinaux provenant de la dernière année du panel, c'est-à-dire 1998, ce qui signifie que seules les

5. Exemple de modélisation de périodes de chômage multiples

5.1 Les données

Dans ce modèle, nous supposons que les fonctions de risque de base et les effets des covariables sont communs aux divers ordres de période. L'ensemble de risques au temps T_j est défini autrement que dans les modèles 1 et 2, et contient toutes les périodes pour lesquelles $t \leq T_j$, en supposant effectivement que toutes les périodes proviennent d'individus différents. Techniquement, ce modèle est un modèle à période unique, de sorte que l'estimation des coefficients et des variances par une méthode robuste sous le plan est simple.

Modèle 3 : Le dernier modèle étudié est le suivant :

$$h_j(t|\mathbf{x}_j) = \lambda_0(t) e^{\mathbf{x}_j(t)\boldsymbol{\beta}}.$$

L'estimation de la matrice de covariance de \mathbf{B} sera faite en utilisant la méthode d'estimation robuste sous le plan expliquée à la section 3.2.

où \mathbf{h}_0^* à la forme de \mathbf{h}_0^* , mais avec $S^{(0)}(t, \mathbf{B})$ et $S^{(1)}(t, \mathbf{B})$ remplacés par $S^{(0)}_j(t, \mathbf{B})$ et $S^{(1)}_j(t, \mathbf{B})$ respectivement.

$$\sum_{k=1}^K \sum_{j=1}^J w_k^*(t) \mathbf{h}_0^*(T_j, \mathbf{B}) = 0,$$

L'estimation sous le plan du paramètre \mathbf{B} s'obtient en remplaçant \mathbf{B} respectivement, mais avec N_j remplaçant n et \mathbf{B} rem-

L'ordre des périodes fait uniquement référence à celles survenant durant la période d'observation pour laquelle les données sont recueillies et non pas à la biographie complète d'un individu (à moins que les deux périodes ne coïncident). Par exemple, par première période, nous entendons la première période de l'état étudié survenant durant la période d'observation, alors qu'il pourrait s'agir d'une période d'ordre absolu plus élevée au cours de la vie de la personne. Cette limite implique qu'il convient d'interpréter avec précaution toute incidence que l'ordre de la période peut avoir sur les effets des covariables ou sur la dépendance par rapport au temps.

Modèle 1 : Dans le premier modèle, l'ensemble de risques est défini avec précaution afin de tenir compte de l'ordre des périodes en ce sens qu'un individu ne peut pas être exposé au risque d'achèvement de la deuxième période avant que la première soit achevée, etc. Ce modèle, connu sous le nom de modèle d'ensemble de risques conditionnels, a été proposé par Prentice, Williams et Peterson (1981) et revu par Lin (1994). Il a également été discuté par Hamerle (1989) et par Blossfeld et Hamerle (1989) dans le contexte de la modélisation de processus à épisodes multiples. En général, l'ensemble de risques conditionnels au temps t pour l'achèvement d'une période d'ordre j comprend tous les individus qui sont dans leur j^{e} période. Ce modèle permet que l'ordre de la période influence à la fois l'effet des covariables et la forme de la fonction de risque de base.

La fonction de risque pour le i^{e} individu pour la période de j^{e} ordre est

$$h_j(t | \mathbf{x}_{ij}(t)) = \lambda_{0j}(t) e^{\mathbf{x}_{ij}(t) \boldsymbol{\beta}_j},$$

où, pour chaque ordre de période, une fonction de risque de base différentielle et pour d'autres que nous examinons à la présente section, le temps t est mesuré à partir du début de la j^{e} période. Bien que les périodes chez un même individu ne soient pas nécessairement indépendantes, la vraisemblance partielle qui suit reste valide pour l'estimation des $\boldsymbol{\beta}_j$:

$$L(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K) = \prod_{N'}^j \prod_{N'}^K \frac{e^{\mathbf{x}_{ij}(T_{ij}) \boldsymbol{\beta}_j}}{\sum_{N'}^j \prod_{N'}^K \frac{e^{\mathbf{x}_{ij}(T_{ij}) \boldsymbol{\beta}_j}}{\delta_{ij}}}. \quad (11)$$

Ici, T_{1j}, \dots, T_{N_jj} sont N_j durées de période de j^{e} ordre éventuellement censurées à droite, $\delta_{ij} = 1$ si T_{ij} est une durée observée et $\delta_{ij} = 0$ autrement, et K est l'ordre le plus élevé des périodes qui doivent être incluses dans le modèle de Cox. Au dénominateur, la somme est calculée sur les j^{e} périodes risquant d'être achevées au temps T_{ij} , c'est-à-dire

$$U_0^j(\mathbf{B}) = \sum_{N'}^j \sum_{N'}^K n_{ij0}(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S_{(1)}(T_{ij}, \mathbf{B}_j)}{S_{(0)}(T_{ij}, \mathbf{B}_j)} \right\}, \quad \text{avec} \quad (12)$$

Les estimations sous le plan des paramètres \mathbf{B}_j sont obtenues en résolvant les équations $\sum_{N'}^j w(s) n_{ij0}^j(T_{ij}, \mathbf{B}_j) = 0$ séparément pour chaque j , où n_{ij0}^j a la forme de n_{ij0}^j , mais avec $S_{(0)}$ et $S_{(1)}$ remplacés par $S_{(0)}^j$ et $S_{(1)}^j$ respectivement. Notons que les poids d'échantillonnage correspondent aux individus et non aux périodes. De même, l'estimation de la matrice de covariance de chaque \mathbf{B}_j se fera séparément, en utilisant la méthode d'estimation robuste sous le plan décrite à la section 2.2. Techniquement, il s'agit d'un ensemble d'analyses distinctes selon l'ordre de la période.

Modèle 2 : Le deuxième modèle étudié est le modèle marginal (Wei, Lin et Weissfeld 1989) :

$$h_j(t | \mathbf{x}_{ij}(t)) = \lambda_{0j}(t) e^{\mathbf{x}_{ij}(t) \boldsymbol{\beta}_j},$$

où, pour chaque ordre de période, nous permettons une fonction de risque de base différentielle tandis que les effets des covariables sont maintenus les mêmes pour différents ordres de période. La fonction de vraisemblance partielle correspondante, ainsi que l'ensemble de risques, sous l'hypothèse que les périodes chez le même individu sont indépendantes, sont les mêmes que pour le modèle 1, avec $\boldsymbol{\beta}_j$ remplaçant le paramètre de population finie est

$$U_0^j(\mathbf{B}) = \sum_{N'}^j \sum_{N'}^K n_{ij0}^j(T_{ij}, \mathbf{B}) = 0, \quad \text{avec}$$

$$n_{ij0}^j(T_{ij}, \mathbf{B}_j) = \delta_{ij} \left\{ \mathbf{x}_{ij}(T_{ij}) - \frac{S_{(1)}(T_{ij}, \mathbf{B}_j)}{S_{(0)}(T_{ij}, \mathbf{B}_j)} \right\},$$

ces conditions, ce qui permet d'obtenir une estimation convergente sous le plan $\mathcal{A}(U_0(\mathbf{b}))$ par application d'une méthode d'estimation de la variance fondée sur le plan à l'expression de rechange, puis l'évaluation de cette estimation de la variance à \mathbf{B} . Si la méthode d'estimation de la variance sous le plan choisie est la méthode de linéarisation, alors la première étape consiste à calculer le résidu suivant pour chacun des individus échantillonnés :

$$(01) \cdot \left\{ \frac{(\mathfrak{g}^{\epsilon_f L})_{(0)} S}{(\mathfrak{g}^{\epsilon_f L})_{(1)} S} - (\epsilon_f L)' \mathbf{x} \right\} \cdot \frac{(\mathfrak{g}^{\epsilon_f L})_{(0)} S}{\mathfrak{g}^{(\epsilon_f L)' \mathbf{x}} \mathfrak{P}^{(\epsilon_f L)' \mathbf{x}} f \mathfrak{g}(s) \omega} \sum_{N=1}^f \frac{1}{N} -$$

Chaque individu compris dans l'échantillon appartient à une UPE particulière dans une strate donnée. Donc, au lieu d'identifier un individu à l'aide d'un indice inférieur unique i , nous utiliserons un indice inférieur triple hci où $h = 1, 2, \dots, H$ identifie la strate, $c = 1, 2, \dots, c_h$ identifie l'UPE dans la strate et $i = 1, 2, \dots, m_{hc}$ identifie l'individu dans l'UPE. Alors,

no

$$\frac{t^{hc}}{\sum_{h_c=1}^I w^{hc} n^{hc}} \text{ et } \underline{t}^{hc} = \frac{1}{\sum_{h_c=1}^C} \frac{(1-t^{hc})c^{hc}}{l} \sum_{H=1}^{l_H} ((\mathfrak{B})^0 \cap) \mathcal{A}_H(\underline{t}^{hc}, \underline{y}_H)$$

3. Inférence pour les modèles du taux de risque à périodes multiples

3.1 Inférence fondée sur le modèle

Si plus d'une période est observée pour un individu, il est raisonnable de supposer que ces périodes ne sont pas indépendantes. Donc, la fonction de vraisemblance partielle (2) est spécifiée incorrectement pour des périodes multiples, puisqu'elle ne tient pas compte de la corrélation intra-individu des périodes observées chez le même individu. En s'inspirant de Lin et Wei (1989), il suffit de modifier seulement la matrice de covariances des paramètres du modèle estimé, puisque les durées corrélées affectent la variance, tandis que les paramètres du modèle peuvent être estimés de manière convergente sans tenir compte de cette corrélation. Lin (1994) démontre comment la matrice de covariance des paramètres du modèle estimé peut être estimée en cas de corrélation intra-individu des périodes, à condition que les périodes provenant d'individus différents soient indépendantes.

4. Trois modèles pour les périodes multiples

compte simplement de la corrélation intra-individu. L'estimation de la variance sous le plan de sondage dans le cas de données hiérarchiques corrélées dans les grappes peut être grandement simplifiée quand il est raisonnable de supposer que les individus provenant d'unités primaires d'échantillonnage (UPÉ) différentes ne sont pas corrélés. Cela équivaut à supposer que les UPÉ sont échantillonnées avec remise. Cette hypothèse est aussi vérifiée approximativement quand les unités de premier degré sont obtenues par échantillonnage sans remise, à condition que la fraction d'échantillonnage au premier degré soit très faible. Dans ces conditions, une estimation de la variabilité inter-UPÉ reflète la variabilité entre les unités à tous les degrés d'échantillonnage subséquents, quelle que soit la structure de dépendance entre les observations dans chaque UPÉ. Pour un cas de données corrélées dans les grappes, voir Williams (2000). Cela implique que l'approche d'estimation robuste de la variance du modèle à période unique dans le cas d'un plan de sondage comportant un échantillonnage avec remise au premier degré décrite par Binder (1992) peut être appliquée directement au cas des périodes multiples, puisqu'elle tient compte de l'effet de la corrélation intra-grappe à tous les niveaux dans chaque UPÉ.

Dans une enquête longitudinale à plan de sondage à plusieurs degrés, les événements multiples peuvent être classés à divers niveaux : les périodes sont regroupées dans un individu et les individus sont groupés dans les unités de dégredé d'échantillonnage élevé. La corrélation intra-grappe positive à tout niveau ajoutée aux estimations calculées à partir de ce genre de données une variation supplémentaire, outre celle attendue sous des conditions d'indépendance. L'hypothèse d'indépendance des observations lorsque celles-ci sont corrélées dans les grappes donne lieu à une sous-estimation des erreurs-types réelles, ce qui exagère les valeurs des statistiques de test et, en dernière analyse, aboutit au rejet trop fréquent des hypothèses nulles. Donc, pour les périodes multiples chez un individu, où les données proviennent d'une enquête longitudinale, il ne suffit pas de tenir compte simplement de la corrélation intra-individu.

Afin de tenir compte de covariables ayant des effets différents pour des périodes d'ordres différents, nous explorons trois modèles pour la définition du risque et les modèles se distinguent par la définition du risque et les hypothèses au sujet du risque de base. Deux de ces modèles tiennent compte de l'ordre des périodes.

score de vraisemblance partielle (3) calculée d'après les périodes de la population finie visée par l'enquête :

$$U^0(\mathbf{B}) = \sum_{i=1}^I n_{i0}^0(T_i, \mathbf{B}) = 0,$$

où $n_{i0}^0(T_i, \mathbf{B})$ est le résidu de score défini de la même manière que $n_{i0}(T_i, \mathbf{B})$, excepté que les moyennes dans les définitions de $S^{(1)}(t, \mathbf{B})$ et de $S^{(1)}(t, \mathbf{B})$ portent sur N plutôt que n observations. Notons que, si les membres de la population finie visée par l'enquête ne connaissent pas tous des périodes de l'état étudié, N représente la taille de la sous-population qui vit de telles périodes et la sommation est faite sur ces N individus.

Une estimation $\hat{\mathbf{B}}$ du paramètre \mathbf{B} est obtenue sous forme d'une solution de l'équation d'estimation du pseudo-score partiel

$$U^0(\hat{\mathbf{B}}) = \sum_{i=1}^I n_i^0(s) \hat{n}_{i0}(T_i, \hat{\mathbf{B}}) = 0,$$

où $n_i^0(s) = w_i$, le poids de sondage, si $i \in s$, et 0 autrement. La fonction $\hat{n}_{i0}(T_i, \hat{\mathbf{B}})$ prend la forme

$$\hat{n}_{i0}(T_i, \hat{\mathbf{B}}) = \delta_i = \left\{ \mathbf{x}_i'(T_i) - \frac{S^{(1)}(T_i, \hat{\mathbf{B}})}{S^{(0)}(T_i, \hat{\mathbf{B}})} \right\},$$

$$S^{(0)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^I n_i^0(s) Y_i^0(t) e^{\mathbf{x}_i'(t) \hat{\mathbf{B}}},$$

$$S^{(1)}(t, \hat{\mathbf{B}}) = \sum_{i=1}^I n_i^0(s) Y_i^1(t) \mathbf{x}_i(t) e^{\mathbf{x}_i'(t) \hat{\mathbf{B}}}.$$

En général, la variance sous le plan de sondage d'une estimation $\hat{\theta}$ qui satisfait une équation d'estimation de la forme $U(\hat{\theta}) = \sum w_i n_i^0(\hat{\theta}) = 0$ peut être estimée, par linéarisation, en tant que

$$J^{-1} V(U(\hat{\theta})) J^{-1}, \quad (9)$$

où $J = \partial U(\hat{\theta}) / \partial \theta$ est évaluée à $\theta = \hat{\theta}$, et $V(U(\hat{\theta}))$ est la variance estimée du total estimé $U(\hat{\theta})$ obtenue par une méthode standard d'estimation de la variance fondée sur le plan (voir, par exemple, Cochran (1977)) et évaluée à $\theta = \hat{\theta}$. Binder (1983) énonce qu'afin d'utiliser cette approche pour calculer une estimation convergente de la variance, $U(\hat{\theta})$ doit être exprimée sous la forme d'une somme de vecteurs aléatoires indépendants. Dans le cas du modèle à risques proportionnels susmentionné, $U^0(\hat{\mathbf{B}})$ ne satisfait pas cette condition, puisque chaque n_{i0}^0 est une fonction de $S^{(0)}(T_i, \hat{\mathbf{B}})$ et de $S^{(1)}(T_i, \hat{\mathbf{B}})$, qui englobent tous deux de nombreux individus outre le i^e . Donc, Binder (1992) a trouvé pour $U^0(\hat{\mathbf{B}})$ une expression de rechange conforme à

ou

$$n_{i0}^0(T_i, \hat{\mathbf{B}}) = \delta_i = \left\{ \mathbf{x}_i'(T_i) - \frac{S^{(0)}(T_i, \hat{\mathbf{B}})}{S^{(1)}(T_i, \hat{\mathbf{B}})} \right\}, \quad (4)$$

et

$$S^{(0)}(t, \hat{\mathbf{B}}) = \frac{1}{I} \sum_{i=1}^I n_i^0(s) Y_i^0(t) e^{\mathbf{x}_i'(t) \hat{\mathbf{B}}}, \quad (5)$$

$$S^{(1)}(t, \hat{\mathbf{B}}) = \frac{1}{I} \sum_{i=1}^I n_i^0(s) Y_i^1(t) \mathbf{x}_i(t) e^{\mathbf{x}_i'(t) \hat{\mathbf{B}}}. \quad (6)$$

Si le modèle (1) est vérifié et que les durées sont indépendantes, la matrice des variances fondées sur le modèle de la fonction de score $U^0(\hat{\mathbf{B}})$ est

$$J(\hat{\mathbf{B}}) = -\partial U^0(\hat{\mathbf{B}}) / \partial \hat{\mathbf{B}}$$

$$= \sum_{i=1}^I \delta_i \left\{ \frac{S^{(2)}(T_i, \hat{\mathbf{B}})}{S^{(1)}(T_i, \hat{\mathbf{B}})} - \frac{[S^{(0)}(T_i, \hat{\mathbf{B}})]^2}{[S^{(1)}(T_i, \hat{\mathbf{B}})]^2} \right\},$$

ou

$$\frac{1}{I} \sum_{i=1}^I Y_i^0(t) \mathbf{x}_i(t) \mathbf{x}_i'(t) e^{\mathbf{x}_i'(t) \hat{\mathbf{B}}}.$$

La variance approximative de $\hat{\mathbf{B}}$, obtenue par linéarisation, est $J^{-1}(\hat{\mathbf{B}})$.

Si la forme de (1) est incorrecte, mais que les observations sont indépendantes, Lin et Wei (1989) donnent l'estimateur robuste de la variance pour $\hat{\mathbf{B}}$ sous la forme

$$J^{-1}(\hat{\mathbf{B}}) G(\hat{\mathbf{B}}) J^{-1}(\hat{\mathbf{B}}), \quad (7)$$

ou

$$G(\hat{\mathbf{B}}) = \sum_{i=1}^I \delta_i^2 G_i(\hat{\mathbf{B}}) G_i(\hat{\mathbf{B}})$$

et

$$G_i(\hat{\mathbf{B}}) = n_{i0}^0(T_i, \hat{\mathbf{B}})$$

$$= \sum_{i=1}^I \delta_i^2 \frac{Y_i^0(T_i) e^{\mathbf{x}_i'(T_i) \hat{\mathbf{B}}}}{S^{(0)}(T_i, \hat{\mathbf{B}})} \left\{ \mathbf{x}_i'(T_i) - \frac{S^{(0)}(T_i, \hat{\mathbf{B}})}{S^{(1)}(T_i, \hat{\mathbf{B}})} \right\}. \quad (8)$$

2.2 Inférence fondée sur le plan de sondage

Pour les observations provenant d'une enquête à plan d'échantillonnage complexe, Binder (1992) a utilisé une méthode de pseudovraisemblance pour estimer les paramètres d'un modèle à risques proportionnels et leurs variances dans le cas d'une seule période par individu. En particulier, il a commencé par définir le paramètre d'intérêt en population finie comme étant la solution de l'équation de

$$h(t) = \lim_{p \rightarrow 0} \text{Prob} \{t \leq T < t + dt | T \geq t\} \cdot \frac{dp}{dt}$$

La valeur de la fonction de risque au temps t est appelée l'intérêt. Les modèles de durée et l'analyse de la durée en général sont formulés et discutés en termes de la fonction de risque et de ses propriétés.

Du point de vue d'un spécialiste du domaine, l'intérêt principal est souvent d'étudier l'effet de certaines covariables sur la distribution de T . Un modèle à risques proportionnels est fréquemment choisi pour ce genre d'étude. Sous le modèle à risques proportionnels, la fonction de risque de la période T , étant donné un vecteur de covariables variant éventuellement en fonction du temps $\mathbf{x}(t) = (x_1(t), \dots, x_p(t))'$ est

$$h(t | \mathbf{x}(t)) = \lambda_0(t) e^{\mathbf{x}(t)'\boldsymbol{\beta}} \quad (1)$$

La fonction $\lambda_0(t)$ est une fonction de risque de base non spécifiée qui donne la forme de $h(t | \mathbf{x}(t))$. Le risque de base décrit la dépendance de la durée, comme préciser si le taux de risque dépend du temps déjà écoulé dans la période. Par exemple, une dépendance négative décrit la situation où la probabilité de sortie d'un état est d'autant plus faible que la période est longue. Si la valeur de toutes les variables $\mathbf{x}(t)$ d'un individu est fixée à 0, la valeur (niveau) de la fonction de risque est égale au risque de base.

2.1 Inférence fondée sur le modèle

Le vecteur $\boldsymbol{\beta}$ contient les paramètres de régression inconnus montrant la dépendance du risque par rapport au vecteur $\mathbf{x}(t)$ et peut être estimé en maximisant la fonction de vraisemblance partielle (Cox 1975) :

$$L(\boldsymbol{\beta}) = \frac{\prod_{i=1}^n \sum_{j=1}^J X_j(T_j) e^{\mathbf{x}_j(T_j)'\boldsymbol{\beta}}}{e^{\mathbf{x}_i(T_i)'\boldsymbol{\beta}}} \quad (2)$$

Ici, T_1, \dots, T_n représentent n durées éventuellement censurées à droite; $\delta_j = 1$ si T_j est une durée observée et $\delta_j = 0$ autrement; et $\mathbf{x}_j(t)$ est le vecteur de covariables correspondant observé sur $[0, T_j]$. Au dénominateur, la somme est calculée sur les périodes qui risquent d'être achevées au temps T_j , c'est-à-dire $X_j = 1$ si $t \leq T_j$, et égale à 0 autrement. L'estimation $\hat{\boldsymbol{\beta}}$ du paramètre $\boldsymbol{\beta}$ du modèle est obtenue en résolvant l'équation de score de vraisemblance partielle

$$U_0(\boldsymbol{\beta}) = \sum_{i=1}^n u_{i0}(T_i, \boldsymbol{\beta}) = 0, \quad (3)$$

nécessaire de tenir compte de l'effet du plan d'échantillonnage sur la distribution des données d'échantillon lors de l'estimation des paramètres du modèle et des variances de ces estimations. Notre approche, lors de l'analyse de données d'enquête complexes, consiste à modéliser les lois marginales des périodes multiples selon des méthodes pour période unique, en traitant la dépendance entre les périodes comme une perturbation - aussi bien la dépendance due à la corrélation des périodes chez la même personne que la dépendance entre individus due au plan de sondage - mais à tenir compte des probabilités de sélection inhérentes à l'aide des poids de sondage. En fonction du modèle choisi, les paramètres de population finie sont définis et estimés comme dans Binder (1992). Les erreurs-types sont estimées par une méthode appropriée de linéarisation convergente sous le plan en posant l'hypothèse que les unités primaires d'échantillonnage sont échantillonnées avec remise dans les strates. Cette hypothèse est valide lorsque les fractions d'échantillonnage de premier degré sont faibles, comme d'échantillonnage le cas dans les enquêtes socioéconomiques. En outre, pour ce genre d'échantillon, la différence entre les inférences en population finie et en superpopulation (c'est-à-dire les erreurs-types et les statistiques de test) s'avère assez négligeable (Lin 2000). Par conséquent, les résultats de l'inférence basée sur notre approche peuvent s'étendre au-delà de la population finie étudiée.

À la section suivante, nous passons en revue la modélisation de périodes uniques et certaines méthodes d'estimation robuste des variances lorsque le modèle est spécifié incorrectement, d'abord dans un cadre fondé sur un modèle, puis dans un cadre fondé sur le plan de sondage. À la section 3, nous discutons plus en détail de l'estimation robuste de la variance pour des périodes multiples. À la section 4, nous introduisons trois modèles pour périodes multiples et décrivons comment les ajuster en utilisant des méthodes d'estimation robuste sous le plan de sondage. À la section 5, nous ajustons ces modèles aux données de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada et discutons des résultats. Enfin, à la section 6, nous présentons certaines remarques générales.

2. Inférence pour le modèle de taux de risque à période unique

La durée d'une période (ou simplement, une période) reçue par un individu est une variable aléatoire dénotée par T . Nous nous intéressons particulièrement à la fonction de risque $h(t)$ de T au temps t , définie comme étant le risque instantané d'achèvement de la période au temps t sachant qu'elle n'a pas été achevée avant le temps t , exprimée formellement par

Modélisation des durées de périodes multiples à partir de données d'enquête longitudinale

Milorad S. Kovačević et Georgia Roberts

Résumé

nos études certaines modifications du modèle de Cox à période unique classiques afin de traiter les périodes multiples et d'échantillonner complexe. Une des modifications essentielles d'une approche fondée sur le plan de sondage pour l'estimation des coefficients est le choix de la variance, dans l'estimation de la variance, chaque individu est traité comme une grappe de périodes, ce qui ajoute un degré supplémentaire de mise en grappes dans le plan de sondage. D'autres modifications sont ajoutées au but de rendre compte la spécification du risque de base afin de tenir compte de la dépendance différentielle éventuelle du risque à l'égard de l'ordre et de la durée des périodes successives, et de tenir compte aussi des effets différentiels des covariables sur les périodes de différents ordres. Ces approches sont illustrées en utilisant des données provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée au Canada.

Mois clés : Régression de Cox; inférence fondée sur le plan de sondage; inférence fondée sur un modèle; ordre de la période; EDTR

1. Introduction

Le problème de modélisation abordé dans le présent article est connu sous divers noms, tels que modélisation des temps de défaillance corrélés, modélisation multivariée de la survie, modélisation de périodes multiples ou problème d'événements récurrents. Il a été étudié dans la littérature biomédicale (par exemple, Lin 1994, Hougaard 1999), sociale (Blossfeld et Hamerle 1989, Hamerle 1988) et économique (Lanoster 1979, Heckman et Singer 1982). Généralement, ce type de modélisation est requis pour résoudre les problèmes qui se posent dans les études du temps écoulé jusqu'à l'événement, lorsque deux événements ou plus surviennent chez le même sujet et que le but de la recherche est d'évaluer l'effet de diverses covariables sur la durée d'une période dans un état particulier. Les temps de défaillance sont caractérisés chez un sujet donné et, donc, l'hypothèse d'indépendance des temps de défaillance conditionnellement à des covariables mesurées que requièrent les modèles de survie standard est vraisemblablement violée. Dans les études de durée des périodes (de pauvreté, de chômage, etc.), la «défaillance» équivaut à la sortie de l'état d'intérêt. Une propriété supplémentaire d'un grand nombre de périodes multiples, souvent ignorée, est que les périodes sont des «événements» ordonnés; autrement dit, la deuxième période ne peut pas survenir avant la première, et ainsi de suite. Le présent article a été motivé par une étude des périodes de chômage dont il est discuté plus en détail à la section 5.

L'interdépendance des périodes survenant chez un même individu est due au fait qu'elles ont en commun certains

Une autre approche, qui est celle que nous utiliserons, consiste à adopter une méthode semi-paramétrique dans laquelle nous ne modélisons pas explicitement l'interdépendance des périodes multiples. Nous modélisons les lois marginales des périodes individuelles, en utilisant éventuellement l'ordre des périodes dans la spécification du modèle. Dans un contexte autre que les sondages, Lin (1994) décrit comment il est suffisant de modifier simplement la matrice de covariance « naïve » des coefficients estimés du modèle obtenue sous l'hypothèse d'indépendance, puisque les durées corrélées doivent être prises en compte dans les estimations de la variance, mais non dans les estimations des coefficients proprement dits.

caractéristiques inobservées de l'individu. L'effet de ces caractéristiques inobservées peut être modélisé explicitement sous forme d'un effet aléatoire (par exemple, Clayton et Cuzick 1985). Le cas échéant, il est supposé que l'effet aléatoire suit une loi statistique connue. La loi gamma de moyenne 1 et de variance inconnue est la loi privilégiée dans de nombreuses applications. Ensuite, des estimations des effets aléatoires et fixes peuvent être obtenues par une méthode appropriée (par exemple, vraisemblance en deux étapes (Lancaster 1979), en utilisant un algorithme EM (Klein 1992), etc.). Cette approche n'est pas explorée dans

$\tau_z = \tau_x$ de sorte que (12) se réduit à (10). En général, $p \leq 1$ et (11) découle de (12). En fait, dans notre appli-cation, nous estimons que p est 0,59, de sorte que, dans (11), les bornes ne devraient pas être très rapprochées.

Remerciements

Les travaux du deuxième auteur ont été financés par la subvention 20.0286/01.3 du Conseil national du Brésil pour le développement scientifique et technologique (CNPq).

Bibliographie

Balaghi, B.H. (2001). *Econometric Analysis of Panel Data*. 2^{ème} Ed. Chichester : John Wiley & Sons, Inc.

Berrington, A. (2002). Exploring relationships between entry into parenthood and gender role attitudes: evidence from the British Household Panel Study. Dans *Meaning and Choice: Value Orientations and Life Course Decisions*. (Ed., R. Lesthaeghe) Brussels : NIDI.

Chambers, R.L., et Skinner, C.J. Eds. (2003). *Analysis of Survey Data*. Chichester : John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-92.

Diggle, P.J., Heagerty, P., Liang, K. et Zeger, S.L. (2002). *Analysis of Longitudinal Data*. 2^{ème} Ed. Oxford : Oxford University Press.

Fan, P.-L., et Marin, M.M. (2000). Influences on gender-role attitudes during the transition to adulthood. *Social Science Research*, 29, 258-283.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Samhva* Vol. 37, Séries C, 117-132.

Fuller, W.A. (1984). Application de la méthode des sondages carrés et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-137.

Fuller, W.A., et Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68, 626-632.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, 74, 430-431.

Goldstein, H. (2003). *Multilevel Statistical Models*, 3^{ème} Ed. London : Arnold.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kish, L., et Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Séries B, 36, 1-37.

Lavange, L.M., Koch, G.G. et Schwartz, T.A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, 20, 2609-23.

Lavange, L.M., Stearns, S., Lataia, J.E., Koch, G.G. et Shah, B.V. (1996). Innovative strategies using SUDAAN for analysis of health surveys with complex samples. *Statistical Methods in Medical Research*, 5, 311-329.

Liang, K.Y., et Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Liang, K.Y., Zeger, S.L. et Qaish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society*, Séries B, 54, 3-40.

Lipsitz, S.R., Fitzmaurice, G.M., Oray, E.J. et Laird, N.M. (1994). Performance of generalized estimating equations in practical situations. *Biometrics*, 50, 270-278.

Lynn, P., et Lieveley, D. (1991). *Drawing General Population Samples in Great Britain*. London : Social and Community Planning Research.

Maas, C.J.M., et Hox, J.J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, 46, 427-440.

Morgan, S.P., et Waite, L.J. (1987). Parenthood and the attitudes of young adults. *Am. Sociological Review*, 52, 541-547.

Pfeiffermann, D., Skinner, C., Holmes, D., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-56.

Renard, D., et Moltenberghs, G. (2002). Multilevel modelling of complex survey data. Dans *Topics in Modelling Clustered Data* (Eds., M. Aerts, H. Geys, G. Moltenberghs et L.M. Ryan). Boca Raton : Chapman and Hall/CRC, 263-272.

Scott, A.J., et Holt, D. (1982). The effect of two stage sampling on ordinary least squares methods. *Journal of the American Statistical Association*, 77, 848-854.

Shah, B.V., Barnwell, B.G. et Bielet, G.S. (1997). SUDAAN User's manual, release 7.5. Research triangle park, NC : Research Triangle Institute.

Skinner, C.J. (1986). Design effects of two stage sampling. *Journal of the Royal Statistical Society*, Séries B, 48, 89-99.

Skinner, C.J. (1989a). Introduction to Part A. In *Analysis of Complex Surveys*, (Eds., C.J. Skinner, D. Holt et T.M.F. Smith) Chichester : Wiley, 23-58.

Skinner, C.J. (1989b). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*. (Eds., C.J. Skinner, D. Holt et T.M.F. Smith) Chichester : Wiley, 59-87.

Skinner, C.J., Holt, D. et Smith, T.M.F. Eds. (1989). *Analysis of Complex Surveys*. Chichester : Wiley.

Comme nous l'avons mentionné à la fin de la section 2, une option est l'utilisation d'une méthode d'estimation de la variance « robuste » basée sur le modèle (5) (Goldstein 2003, page 80). Les valeurs de ces estimations robustes de l'erreur-type sont également incluses dans le tableau 4. Comme nous l'avons prévu à la section 2, l'estimateur robuste de l'erreur-type pour le modèle à deux niveaux donne des résultats très semblables à ceux de l'estimateur par linéarisation qui ne tient pas compte de la mise en grappes. L'estimateur robuste de l'erreur-type pour le modèle à trois niveaux donne des résultats fort semblables à ceux de l'estimateur par linéarisation tenant compte de l'échantillonnage à deux degrés. Les légers écarts reflètent les différences entre les méthodes d'estimation de V .

La méthode de linéarisation en présence d'échantillonnage à deux degrés s'approche donc fort de la méthode d'estimation robuste de la variance utilisée dans la littérature

Tableau 4 Erreurs-types estimées des coefficients de régression

Linéarisation		Modélisation multiniveaux			
EAS Complexe	Fondée sur le modèle à 2 niveaux	Robuste à 2 niveaux	Fondée sur le modèle à 3 niveaux	Robuste à 2 niveaux	Robuste à 2 niveaux
Ordonnée à l'origine	0,287	0,253	0,288	0,259	0,293
Année, t	0,014	0,013	0,014	0,013	0,014
Groupe d'âge					
16 à 21 ans	0,191	0,155	0,192	0,155	0,243
22 à 27 ans	0,214	0,187	0,215	0,187	0,266
28 à 33 ans	0,237	0,218	0,238	0,218	0,271
34 ans et plus					
Activité économique					
Occupée temps plein	0,103	0,098	0,103	0,098	0,096
Autre inactive	0,166	0,150	0,166	0,146	0,148
Étudiante temps plein	0,207	0,238	0,207	0,199	0,236
Soin de la famille	0,125	0,102	0,112	0,112	0,101
Diplôme					
Qualif.	0,228	0,210	0,207	0,228	0,211
Niveau A	0,238	0,239	0,209	0,240	0,237
Niveau O	0,234	0,199	0,217	0,235	0,199
Autre	0,247	0,224	0,229	0,249	0,223

L'observation de Kish et Frankel (1974), à savoir que les valeurs du meff n'ont pas tendance à être plus élevées pour les coefficients de régression que pour les moyennes de la variable dépendante.

Tableau 2 Estimations pour la régression avec des covariables définies selon l'activité économique

Vagues	β e-t.		meff	
	1-9	1-9	1,3	1,3,5 1-7
Ordonnée à l'origine	20,58	0,11	1,13	1,01 1,09 1,38 1,50
Contrastes pour				
Occupés temps plein	-1,03	0,10	0,93	0,91 0,89
Autre inactive	-0,80	0,15	0,60	0,96 0,68 0,76 0,81
Étudiante temps plein	0,41	0,24	1,10	1,32 1,14 1,48 1,44
Soin de la famille	-2,18	0,10	0,72	0,49 0,58 0,66 0,60
Nota : a) L'ordonnée à l'origine est la moyenne pour les femmes occupées à temps plein				
b) Les contrastes sont calculés pour les autres catégories d'activité économique relativement au travail à temps plein.				

Tableau 3 Estimations pour les coefficients de régression avec covariables supplémentaires dans le modèle

Vagues	β e-t.		Meff	
	1-9	1-9	1	1,3 1,3,5 1-7 1-9
Ordonnée à l'origine	20,20	0,30	0,95	0,87 1,04 1,07
Année, t				
	-0,04	0,01	-	0,86 0,69 0,59 0,96
Groupe d'âge				
16 à 21 ans	0,00	-		
22 à 27 ans	-0,71	0,25	1,22	1,37 1,44 1,73 1,64
28 à 33 ans	-0,89	0,27	1,38	1,40 1,46 1,68 1,59
34 ans et plus	-1,03	0,27	0,94	1,10 1,13 1,26 1,34
Activité économique				
Occupés temps plein	0,00	-	0,97	0,95 0,96 1,06 0,91
Autre inactive	-0,75	0,15	0,60	0,96 0,68 0,77 0,81
Étudiante temps plein	0,17	0,24	0,93	1,32 1,23 1,39 1,32
Soin de la famille	-2,09	0,10	0,77	0,59 0,70 0,78 0,67
Qualification				
Diplôme	0,00	-	0,77	0,64 0,75 0,87 0,85
Qualif.	-0,52	0,21	0,98	0,87 0,94 0,94 1,01
Niveau A	-0,61	0,24	0,62	0,62 0,59 0,69 0,73
Niveau O	-0,44	0,20	0,83	0,83 0,78 0,80 0,82
Autre	-1,16	0,22	0,83	0,83 0,78 0,80 0,82

Nous considérons ensuite les erreurs-types fondées sur le modèle obtenues à partir du modèle à trois niveaux (5), comme il est discuté à la section 2. Les résultats sont présentés au tableau 4 dans la colonne intitulée « fondée sur le modèle à trois niveaux ». Aux fins de comparaison, nous estimons aussi les erreurs-types sous le modèle à deux niveaux (2) et présentons les résultats dans la colonne intitulée « fondée sur le modèle à deux niveaux ». Les estimations qui figurent dans ces deux colonnes sont presque identiques. Il existe un écart d'un chiffre au niveau de la troisième décimale pour certains coefficients et un écart un peu plus important pour l'ordonnée à l'origine. Nous pensons qu'il s'agit d'une preuve qu'ajouter simplement un terme d'effet régional aléatoire peut donner lieu à une sous-estimation importante de l'effet de la mise en grappes sur les erreurs-types estimées des coefficients de régression. Ces données sont en accord avec la borne supérieure théorique du meff donnée en (11). La valeur estimée de τ dans l'expression (11) est 0,019 et aucune des covariables ne devrait, en principe, présenter une corrélation intra-régionale importante, de sorte que les valeurs prévues des estimateurs de la variance pour les modèles à deux et à trois niveaux devraient être très proches.

Nous avançons à la section 3 que la caractéristique de la mise en grappes principalement susceptible d'avoir une incidence sur la matrice de covariance de β est la variation inter-grappes des coefficients de régression. Nous avons exploré cette idée en introduisant des coefficients aléatoires dans le modèle. En traitant alors les éléments de β comme les valeurs prévues des coefficients aléatoires, nous avons constaté que les estimations de β avaient à peine changé. Nous avons trouvé que les erreurs-types estimées de ces estimations étaient, en effet, exagérées, et ce, bien davantage que par introduction de l'effet aléatoire supplémentaire de grappe dans le modèle (5), et que l'accroissement était du même ordre de grandeur que ceux des meffs des tableaux 2 et 3. Néanmoins, la méthode des MCGI a produit plusieurs estimations négatives des coefficients aléatoires qu'il faut laisser varier ou, plus généralement, sur la spécification du modèle. Ce problème s'accroît à mesure qu'augmente le nombre de covariables, car le nombre de paramètres dans la matrice de covariance du vecteur de coefficients augmente en fonction du carré du nombre de covariables. Dans l'ensemble, l'introduction de coefficients aléatoires semble créer au moins autant de problèmes qu'elle n'en résout si la mise en grappes ne présente pas d'intérêt scientifique intrinsèque et ne semble pas être un moyen très satisfaisant de tenir compte de la mise en grappes dans l'estimation de la variance. Il est plus simple de changer de méthode d'estimation.

Nous étions ensuite l'analyse en introduisant des variables indicatrices d'activité économique comme co-variables. Le modèle de régression résultant comprend un terme d'ordonnée à l'origine et quatre covariables représentant les contrastes entre les femmes occupées à temps plein et celles appartenant à d'autres catégories d'activité économique. Les valeurs estimées du meff sont présentées au tableau 2. L'ordonnée à l'origine est une moyenne de domaine et, selon la théorie classique du meff, d'une valeur de 1,51 du tableau 1. Comme auparavant, le tableau 2 semble indiquer une tendance du meff à augmenter, de 1,13 dans le cas d'une seule vague à 1,50 dans le cas de cinq vagues, quoique ces valeurs soient plus faibles qu'au tableau 1. La taille du meff pour les contrastes du tableau 2 varie, certaines valeurs étant supérieures et d'autres inférieures à un. Ces valeurs du meff peuvent être considérées comme une combinaison de l'effet classique d'augmentation de la variance due à la mise en grappes dans les enquêtes et de l'effet de réduction de la variance due à la mise en blocs dans une expérience. Cette réduction de la variance a lieu si les domaines comparés ont un effet de grappe commun (de la forme η_0 dans le modèle (5)) qui a tendance à s'annuler dans les contrastes, ce qui sous-entend que la variance réelle du contraste est plus faible que l'espérance de l'estimateur de variance fondé sur l'hypothèse d'indépendance entre domaines. Cette dernière espérance sera augmentée par les effets communs. La caractéristique de ces résultats qui présente le plus d'intérêt ici est que, de nouveau, le meff n'a pas tendance à converger vers l'unité à mesure que le nombre de vagues augmente. Si tant est qu'il y ait une tendance, celle-ci est de direction opposée. Pour le contraste d'intérêt scientifique ici, c'est-à-dire celui entre les femmes occupées à temps plein et celles qui « restent à la maison pour s'occuper de la famille », le meff est systématiquement nettement inférieur à un.

Nous perfectionnons ensuite davantage le modèle en incluant, comme covariables supplémentaires, le groupe d'âge, l'année et les qualifications. Les valeurs estimées du meff sont données au tableau 3. Le meff pour les coefficients de régression correspondant aux catégories d'activité économique varie de nouveau, certaines valeurs étant supérieures et d'autres inférieures à l'unité, pour les mêmes raisons que pour les contrastes (qui pourraient également être interprétés comme des coefficients de régression) du tableau 2. De nouveau, il semble que le meff ait tendance à s'écarter de l'unité à mesure que le nombre de vagues augmente. Une comparaison des tableaux 1 et 3 confirme

Pour évaluer l'effet de l'aspect longitudinal des données, nous estimons une série de meffs en utilisant les données provenant des vagues 1, ..., t pour $t = 2, 3, \dots, 5$. Bien que ces meffs estimés soient sujets à une erreur d'échantillonnage, il semble évident, si l'on examine le tableau 1, que le meff tend à augmenter avec le nombre de vagues. Cette tendance pourrait être escamotée, compte tenu de la discussion théorique de la section 3, si le niveau moyen des attitudes égalitaires dans une région varie moins d'année en année que les cotes d'attitude individuelles des femmes. Cela paraît vraisemblable, puisque les secondes seront affectées à la fois par l'erreur de mesure et par les changements réels d'attitude, si bien qu'on pourrait prévoir que $\text{var}(\eta_a)$ diminue plus lentement avec T que $\text{var}(\eta_a) + v_a$. Nous pouvons donc nous attendre à ce que τ_1 , et par conséquent le meff, augmente à mesure que T croît, comme nous l'observons au tableau 1.

β	e-t	meff
19,83	0,12	1,51
1,9	1,9	1,50
1,3	1,3	1,68
1,7	1,7	1,81
1,84		

Tableau 1 Estimations pour les moyennes longitudinales

Nous commençons par estimer le meff pour l'estimateur par la linéarisation, comme il est discuté au début de la section 3. À l'aide de données provenant uniquement de la première vague et en fixant $x_{\text{all}}^{\text{meff}} = 1$, le meff estimé pour cette moyenne transversale donné dans le tableau 1 est de l'ordre de 1,5. Cette valeur est plausible car, si nous procédons à l'approximation habituelle de (9) pour des tailles d'échantillon de grappe inégales en remplaçant m par m , la taille moyenne d'échantillon par grappe, nous trouvons que $1 + (m - 1)\tau = 1,5$ et $m = 1,340/47 \approx 29$ impliquent une valeur de τ d'environ 0,02 et cette faible valeur est en accord avec d'autres valeurs estimées de τ obtenues pour des variables attitudinales dans les enquêtes britanniques (Lynn et Lienesley 1991, annexe D).

Nous commençons par estimer le meff pour l'estimateur par la linéarisation, comme il est discuté au début de la section 3. À l'aide de données provenant uniquement de la première vague et en fixant $x_{\text{all}}^{\text{meff}} = 1$, le meff estimé pour cette moyenne transversale donné dans le tableau 1 est de l'ordre de 1,5. Cette valeur est plausible car, si nous procédons à l'approximation habituelle de (9) pour des tailles d'échantillon de grappe inégales en remplaçant m par m , la taille moyenne d'échantillon par grappe, nous trouvons que $1 + (m - 1)\tau = 1,5$ et $m = 1,340/47 \approx 29$ impliquent une valeur de τ d'environ 0,02 et cette faible valeur est en accord avec d'autres valeurs estimées de τ obtenues pour des variables attitudinales dans les enquêtes britanniques (Lynn et Lienesley 1991, annexe D).

4. Exemple : analyse par la régression des données de la BHPS sur les attitudes à l'égard des rôles de l'homme et de la femme

Nous présentons maintenant une application aux données de la BHPS afin d'illustrer certaines propriétés théoriques discutées à la section précédente.

Ces dernières décennies, nous avons été témoins d'une évolution importante des rôles de l'homme et de la femme au sein de la famille dans de nombreux pays. Les spécialistes des sciences sociales s'intéressent à la relation entre l'évolution des attitudes à l'égard des rôles de l'homme et de la femme, d'une part, et les changements de comportement, la forme, d'autre part, à la condition parentale et à la participation au marché du travail, d'autre part (par exemple, Morgan et Waite 1987; Fan et Martin 2000). Diverses formes d'analyse statistique sont utilisées pour fournir des preuves de ces relations. Ici, nous considérons l'estimation d'un modèle linéaire de la forme (1) concernant l'égard variable de résultat, y , une mesure de l'attitude à l'égard des rôles de l'homme et de la femme, à la suite d'une analyse de Berthoin (2002).

Les données proviennent des vagues 1, 3, 5, 7 et 9 (recueillies en 1991, 1993, 1995, 1997 et 1999, respectivement) de la BHPS et ces vagues sont codées $t = 1, \dots, T = 5$ respectivement. On a demandé aux répondants s'ils étaient « tout à fait d'accord », « d'accord », « ni d'accord ni en désaccord », « en désaccord » ou « tout à fait en désaccord » avec une série d'énoncés concernant la famille, ainsi que les rôles de la femme et le travail à l'extérieur du ménage. Les réponses ont été cotées de 1 à 5. On a recouru à l'analyse factorielle pour évaluer quels énoncés pourraient être combinés en une mesure de l'attitude à l'égard des rôles de l'homme et de la femme. La cote d'attitude, y_t , consistait ici est la cote totale pour les six énoncés choisis pour la femme t lors de la vague t . Plus la cote est élevée, plus les attitudes à l'égard des rôles de l'homme et de la femme sont égalitaires. Berthoin (2002) présente une discussion détaillée de cette variable. Une analyse plus complexe pourrait inclure un modèle d'erreur de mesure pour les attitudes (par exemple, Fan et Martin 2000), chacune des réponses sur l'échelle de cinq points aux six énoncés étant traitée comme une variable ordinaire. Ici, nous adoptons une approche plus simple consistant à traiter la cote agrégée y_t comme le vecteur de coefficients comme β comme présentant un intérêt scientifique et à inclure l'erreur de mesure dans le terme d'erreur du modèle.

Nous nous sommes fondés sur la discussion de Berthoin (2002) en vue de choisir les covariables pour l'analyse par la régression, mais avons réduit leur nombre afin de nous concentrer plus facilement sur les questions

méthodologiques d'intérêt. La covariable présentant le principal intérêt scientifique est l'activité économique, qui permet de faire la distinction, en particulier, entre les femmes qui restent au foyer pour s'occuper des enfants (dénote « soin de la famille ») et celles qui poursuivent d'autres formes d'activité reliée au marché du travail. Les variables reflétant l'âge et le niveau de scolarité sont également incluses, puisqu'il a été souvent démontré qu'elles sont fortement corrélées aux attitudes à l'égard des rôles de l'homme et de la femme (par exemple, Fan et Martin 2000). Les valeurs de toutes ces covariables peuvent varier d'une vague à l'autre de l'enquête. Une variable d'année (prenant les valeurs 1, 3, ..., 9) est également incluse. Elle peut refléter à la fois les changements chronologiques et le vieillissement général des femmes comprises dans l'échantillon.

La BHPS est une enquête-ménage par panel réalisée auprès des membres de la population à domicile de la Grande-Bretagne (Taylor, Brice, Buck et Prentice-Lane 2001). L'échantillon initial (vague 1) a été sélectionné en 1991 selon un plan stratifié à plusieurs degrés dans lequel les probabilités d'inclusion des ménages étaient à peu près égales. Les ménages ont été regroupés en 250 unités primaires d'échantillonnage (UPB) correspondant aux secteurs postaux. Tous les membres résidents de 16 ans et plus ont été sélectionnés dans les ménages échantillonnés. Tous les adultes sélectionnés durant la première vague ont été suivis lors de la deuxième vague et ainsi de suite, et représentent le panel longitudinal. L'enquête est sujette à une érosion de l'échantillon et à d'autres formes de non-réponse à une vague. Pour traiter cette non-réponse, nous avons simplement remplacé s dans (3) par l'« échantillon longitudinal » d'individus pour lesquels les observations étaient disponibles pour chacune des vagues $t = 1, \dots, T$ et nous avons choisi de n'appliquer aucun poids de sondage, puisque notre but est d'étudier les effets d'erreur de spécification éventuellement associés à la mise en grappes et nous voulons éviter de confondre ces effets avec ceux de la pondération. Nous ignorons également l'effet de la stratification dans le travail numéroté de la présente section (mais nous présentons à la section 5 certains commentaires sur l'effet de la pondération et de la stratification).

Puisque nous nous intéressons dans l'analyse à la question de savoir si l'activité principale des femmes consiste à « prendre soin de la famille », nous définissons notre population étudiée comme étant les femmes de 16 à 39 ans en 1991. Donc, nos données correspondent à l'échantillon longitudinal de femmes dont l'âge se situe dans la fourchette admissible et ayant répondu à toutes les questions de l'interview (enregistrément complet) lors de chacune des cinq vagues, ce qui donne un échantillon de $n = 1\,340$ femmes. Ces femmes sont réparties de manière relativement

basé sur le modèle à deux niveaux (2) est :

$$\text{meff} = 1 + (m - 1)\tau_1\tau_x \quad (10)$$

ou $\tau_1 = \sigma_{\eta}^2 / (\sigma_{\eta}^2 + \sigma_{\epsilon}^2)$ et τ_x est la corrélation intra-grappe pour x (Scott et Holt 1982; Skinner 1989b, page 68). Ce résultat s'étend, dans le cas longitudinal, à :

$$1 \leq \text{meff} \leq 1 + (m - 1)\tau_1\tau_z \quad (11)$$

où τ est la version de long terme ($T = \infty$) de τ (voir

l'annexe) et τ_z est un coefficient de corrélation intra-grappe pour $z_{it} = \sum_{t=1}^T x_{it}^{meff} / T$. La preuve de ce résultat et les hypothèses simplificatrices requises sont esquissées à l'annexe. Le point principal est que τ et τ_z seront souvent faibles, auquel cas τ_z sera très faible, et donc, la valeur de meff pourrait être invraisemblablement proche de un, l'estimateur de la variance basé sur le modèle présentant un biais par défaut. Nous explorons cet aspect empiriquement à la section 4. Naturellement, nous pourrions introduire des coefficients aléatoires dans le modèle (5), et nous examinons cela également à la section 4. Toutefois, étant donné la difficulté qu'il y a à spécifier correctement un modèle à coefficients aléatoires, il semble peu probable que cette approche soit très robuste.

À la présente section, nous nous sommes concentrés jusqu'à présent sur le biais (ou non-convergence) éventuel des méthodes d'estimation de la variance. Mais il est également souhaitable d'examiner leur efficacité. En particulier, nous pourrions nous attendre à ce que la méthode de linéarisation soit moins efficace que l'estimation de la variance basée sur le modèle, si ce dernier est correct. En principe, l'importance relative de l'efficacité par rapport au biais devrait augmenter à mesure que le nombre de grappes diminue. Wolter (1985, chapitre 8) résume un certain nombre d'études par simulation conçues pour examiner le biais et la variance de l'estimateur de variance par linéarisation qui laissent entendre que la méthode de linéarisation donne de bons résultats même si le nombre de grappes est faible. Fuller (1984) discute de corrections éventuelles des intervalles de confiance en fonction du nombre de degrés de liberté pour les coefficients de régression basés sur la méthode de linéarisation lorsque le nombre de grappes est faible. Une étude par simulation des estimateurs pour les modèles multiniveaux décrite dans Maas et Hox (2004) ne permet pas de conclure que la méthode de linéarisation donne de nettement moins bons résultats que l'approche basée sur une modèle, en ce qui a trait à la couverture des intervalles de confiance pour les coefficients de β , même si le nombre de grappes est aussi faible que 30.

En fait, l'équation (9) est encore vérifiée si nous remplaçons η_a par un effet variable en fonction du

temps η_{at} à condition que nous remplaçons τ par $\tau = \text{var}(\eta_a) / [\text{var}(\eta_a) + \sigma_{\epsilon}^2] / T$, ou $\eta_a = \sum_{t=1}^T \eta_{at} / T$. Dans ces conditions, le meff augmentera à mesure que T augmente si (et uniquement si) $\sigma_{\eta}^2 + \sigma_{\epsilon}^2 / T$ diminue plus rapidement avec T que $\text{var}(\eta_a)$. Qu'il en soit ainsi ou non dépend de l'application particulière. Cependant, nous soutenons que, pour de nombreuses enquêtes longitudinales auprès d'individus avec les grappes fondées sur les régions

(le genre de contexte auquel nous pensons), cette condition est plausible. Dans de telles applications, nous pouvons souvent nous attendre à ce que la valeur de σ_{η}^2 soit grande relativement à celle de σ_{ϵ}^2 (c'est-à-dire que la corrélation intra-grappe transversale soit faible), en particulier à cause d'une erreur de mesure propre à la vague d'enquête et, donc, à ce que $\sigma_{\eta}^2 + \sigma_{\epsilon}^2 / T$ diminue assez rapidement à mesure que T augmente. En principe, les caractéristiques socio-économiques des régions sont souvent plus stables et dans des situations inhérentes seulement de variabilité attendue à ce qu'une erreur de mesure donne lieu à une variance propre à la vague d'enquête importante dans η_{at} . Donc, selon nous, dans de telles applications, on pourrait habituellement s'attendre à ce que le ratio de $\text{var}(\eta_a)$ pour $T = 5$, disons, comparativement à $T = 1$ soit plus grand que $(\sigma_{\eta}^2 + \sigma_{\eta}^2 / 5) / (\sigma_{\eta}^2 + \sigma_{\epsilon}^2)$ qui s'approchera de 1/5 à mesure que $\sigma_{\eta}^2 / \sigma_{\epsilon}^2$ tend vers 0. Nous sommes donc d'avis que, dans de

Nous présentons un exemple empirique à la section 4.

Considérons maintenant les propriétés des estimateurs de variance basés sur le modèle à trois niveaux (5). Nous examinons que l'approche fondée sur l'hypothèse d'effets aléatoires homoscedastiques suivant une loi normale, en ignorant les poids de sondage, étant donné l'équivalence (virtuelle) de l'approche multiniveaux « robuste » et de la

linéarisation. Si le modèle (5) est correct et que nous ignorons effectivement les poids de sondage, alors l'estimateur de la variance basé sur le modèle sera convergent (Goldstein 1986). Cependant, comme il est discuté dans Skinner (1989b, page 68) et corroboré par la théorie dans Skinner (1986), la principale caractéristique de la mise en grappes susceptible d'avoir une incidence sur les erreurs-types des coefficients de régression est la variation inter-grappes des coefficients de régression, ce qui n'est pas pris en compte dans le modèle (5).

Afin de voir comment le modèle (5) pourrait ne pas refléter correctement les effets de la mise en grappes, considérons le cas transversal ($T = 1$), où x est un scalaire. Alors, si le modèle à trois niveaux (5) tient, une expression

homoscédastiques normaux, nous pouvons employer une méthode d'estimation « robuste » de la variance (Goldstein 2003, page 80). Cette approche est étendue à l'utilisation des poids de sondage dans Pfeffermann et coll. (1998). À part la stratification, l'estimateur de la variance est identique à l'estimateur par linéarisation (4) pour une valeur donnée de \hat{p} .

3. Propriétés des estimateurs de variance

À la présente section, nous considérons les propriétés des estimateurs de la matrice de covariance de β décrite à la section précédente. Nous examinons d'abord l'estimateur par itération $v(\beta)$ donné par (4).

La convergence de $v(\beta)$ pour la matrice de covariancé de β suit les arguments établis dans un cadre asymptotique approprié (par exemple, Fuller 1975; Binder 1983). La seule caractéristique non standard est la présence de A^{-1} dans $v(\beta)$ et la dépendance de A à l'égard de β . En fait, dans les grands échantillons, la matrice de covariancé de β ne dépend de β que par la voie de sa valeur limite p^* (dans un

cadre asymptotique donné). Pour le voir, écrivons $\beta - \beta = (\sum_{i=1}^n u_i' A^{-1} z_i, \text{ où } u_i = w_i' x_i' A^{-1} z_i, \text{ et } z_i = y_i - \beta x_i$. Notons que, sous des conditions de régularité faibles (Fuller et Battese 1973, corollaire 3), la distribution asymptotique de $\hat{\beta} - \beta$ est la même que celle de $\hat{\beta} - \beta = (\sum_{i=1}^n u_i' A^{-1} z_i, \text{ où } u_i = w_i' x_i' A^{-1} z_i, \text{ ou } u_i = w_i' x_i' A^{-1} z_i$, et $\hat{\beta}$ prend la même forme que A avec $\hat{\beta}$ remplacé par $p = d \lim(p)$, la limite de probabilité de $\hat{\beta}$ dans le cadre

asymptotique. En écrivant $\bar{z} = \sum z_i/n$ et $\bar{U} = \sum U_i/n$, nous pouvons donc approximer la matrice de covariance de β asymptotiquement par $\text{var}(\beta) \approx (U^{-1} - \bar{U}^{-1}) \text{var}(\bar{z}) U^{-1}$. Si le modèle de travail (2) tient, alors $p = p$ et cette matrice de covariance sera la même pour toute méthode convergente d'estimation de p . Même si le modèle de travail ne tient pas, $\text{var}(\beta)$ sera convergent pour $U^{-1} \text{var}(\bar{z}) U^{-1}$ dans le genre de cadre asymptotique considéré par Fuller (1975) et Binder (1983), ainsi que sous les genres de conditions de régularité que ces auteurs et Fuller et Battese (1973) établissent.

Ensuite, nous examinons l'effet sur la méthode de

linéarisation de la non-prise en compte du plan d'échantillonnage complexe. Nous désignons par $v_0(\beta)$ l'estimateur par linéarisation obtenu d'après l'expression (4) en ignorant le plan, c'est-à-dire en supposant qu'il n'existe qu'une seule strate contenant des UPF identiques aux individus, de sorte que $m_n = n$ est la taille globale d'échantillon et que z_{n0} remplacé par $z' = w_1'x_1'w_1^{-1}e_1$. Nous nous intéressons aux biais de $v_0(\beta)$ quand le plan de sondage est, en fait, complexe. Soit β_k le k^e élément de β et soit $v_0(\beta_k)$ le k^e élément de $v_0(\beta)$. Alors, à l'exemple de Skinner (1989a, page 24), nous mesurons le biais relatif de l'estimateur de

ment relié à la notion d'effet de plan. Pour étudier la nature de $\text{meff}[\hat{\beta}_k, v_0(\hat{\beta}_k)]$, nous commençons par écrire :

$$[(1-u)/u]_{-}({}^1n^s\mathfrak{Z}) = (\mathfrak{g})^0\Lambda$$

$$(9) \quad {}_1^{\prime} \left({}_1^n \overline{\Sigma} \right) [{}_1^{\prime} (\underline{z} - {}_1^{\prime} z) (\underline{z} - {}_1^{\prime} z)^s \overline{\Sigma}] \times$$

tique, nous avons $E[\mathbf{v}_0^{(j)}] \approx \mathbf{U}_{-1}^{-1} \mathbf{n}_1^{(j)} \mathbf{U}_{-1}^{-1}$, où $\mathbf{S}_z^{(j)}$ est la matrice de covariance en limite de probabilité de $\mathbf{z}_j^{(j)}$. Étant donné que le numérateur de $\text{meff}[\mathbf{f}(\mathbf{y})^{(j)} \mathbf{v}_0^{(j)}]$ peut être approximé par $\mathbf{U}_{-1}^{-1} \text{var}(\mathbf{z}_j^{(j)})$, nous pouvons écrire :

$$(L) \quad \frac{{}^{\chi}(\underline{1})[{}^z S_{1-u}]^{\chi}(\underline{1})}{{}^{\chi}(\underline{1})(\underline{z})\text{var}^{\chi}(\underline{1})} = [({}^{\chi}\mathfrak{g})^0 \wedge {}^{\chi}\mathfrak{g}]_{\text{eff}}$$

où $(\underline{U}^{-1})^k$ est la k^{e} ligne de \underline{U}^{-1} . Si $q = 1$, cette expression se simplifie en :

$$(8) \quad [\cdot, \cdot]_{S_{1-n}}^z / (\cdot, \cdot)_{\text{var}} = [(\cdot)^0, \cdot]_{\text{eff}}$$

Nous pouvons explorer des formes plus spécifiques de ces expressions sous divers modèles et hypothèses au sujet des pondérations et du plan d'échantillonnage. Nous nous concentrons ici sur l'effet de la mise en grappes, en supposant que les pondérations sont égales et qu'il n'existe pas de stratification. Considérons le modèle à trois niveaux suivants (5) et, pour simplifier les choses, supposons que $q = 1$ et $x_{ind}^{ind} = 1$ et que β est la moyenne de y^{ind} . Alors, des calculs algébriques simples montrent que la valeur de β est la même pour toutes les grappes.

du i dans la grappe a est $[1 + p \cdot (T - 1)]^{-1}$.

(6) $\sum_{i=1}^n (n_i)^{1/p} = n + n^{1/p} \text{ merr}[\beta, \nu, 0] (1 - m) + 1$

$$(6) \quad {}^1(1 - m) + 1 = [(\mathfrak{g})^0, \mathfrak{g}]_{\text{eff}}$$

où $1 = \sigma^2_{\epsilon} / (\sigma^2_{\epsilon} + \sigma^2_{\eta} + \sigma^2_{\tau})$ est la corrélation intra-grappe de z_i . Nous voyons que, sous ce modèle, le meilleur estimateur de la variance est plus important dans le cas longitudinal que dans le problème transversal (où $T = 1$). Cette constatation dépend de l'hypothèse assez forte selon laquelle les effets de grappe η_a sont constants a

d'éditions, ou vagues, $t = 1, \dots, T$ d'une enquête. Nous parlerons d'individus pour les unités, quoique notre discussion soit d'application plus générale. Soit y_{it}^u la valeur d'une variable résultat pour l'individu $i \in U$ à la vague t et soit $y_i = (y_{i1}^u, \dots, y_{iT}^u)$ le vecteur de mesures répétées. Soit x_{it}^u un vecteur $1 \times q$ correspondant de valeurs des covariables pour l'individu i à la vague t et soit $x_i = (x_{i1}^u, \dots, x_{iT}^u)$. Nous supposons que le modèle linéaire qui suit est vérifié pour l'espérance de y_i sachant $(x_{i1}^u, \dots, x_{iT}^u)$:

$$(1) \quad E(y_i) = x_i \beta,$$

où β est un vecteur $q \times 1$ de coefficients de régression et l'espérance est calculée sous le modèle. Nous supposons que β est la cible de l'inférence, autrement dit que les coefficients de régression sont les paramètres qui intéressent principalement l'analyste. Nous examinerons également d'autres caractéristiques du modèle, comme la matrice de covariance de y_i , mais nous supposons qu'elles sont d'un intérêt secondaire pour l'analyste.

Les données disponibles pour l'inférence au sujet de β proviennent d'une enquête longitudinale dans laquelle les valeurs de y_{it}^u et de x_{it}^u sont observées lors de chaque édition (vague) $t = 1, \dots, T$ pour les individus i dans un échantillon, s , tiré à partir de U à la vague 1 selon un plan d'échantillonnage spécifié. Pour simplifier, nous supposons ici qu'il n'y a pas de non-réponse, mais nous examinons cette possibilité à la section 4.

Afin de formuler un estimateur ponctuel de β , nous étendons la spécification de (1) au modèle « de travail » suivant :

$$(2) \quad y_{it}^u = x_{it}^u \beta + u_{it} + v_{it},$$

où u_{it} et v_{it} sont des effets aléatoires indépendants de moyenne nulle et de variances $\sigma_u^2 = \rho \sigma^2$ et $\sigma_v^2 = (1 - \rho) \sigma^2$ respectivement, sachant $(x_{i1}^u, \dots, x_{iT}^u)$. Ce modèle peut être appelé un modèle de corrélation uniforme (Diggle et coll. 2002, page 55) ou un modèle à deux niveaux (Goldstein 2003). Le paramètre ρ est la corrélation intra-individu.

L'estimateur ponctuel de base de β que nous considérons est

$$(3) \quad \hat{\beta} = \left(\sum_{i \in s} w_i x_i' A_i^{-1} x_i \right)^{-1} \sum_{i \in s} w_i x_i' A_i^{-1} y_i,$$

où w_i est un poids de sondage et A_i est une matrice de covariance $T \times T$ estimée de y_i sous le modèle de travail (2), c'est-à-dire qu'elle contient des éléments diagonaux $\hat{\sigma}_i^2$ et des éléments non diagonaux $\hat{\rho}_i \hat{\sigma}_i^2$, où $(\hat{\rho}_i, \hat{\sigma}_i^2)$ est un estimateur de (ρ, σ^2) . (Notons qu'en fait, $\hat{\sigma}_i^2$ s'annule dans (3) et que σ^2 ne doit donc pas être estimé pour $\hat{\beta}$). En l'absence des termes de pondération et des considérations

étendons les travaux de ces auteurs à la situation où l'on obtient des observations d'enquête longitudinale basées sur un échantillon initial tiré selon un plan d'échantillonnage complexe, en nous concentrant de nouveau sur le cas d'un plan de sondage en grappes. Nous considérons pour ces données longitudinales une classe standard de modèles de régression linéaire qui est décrite dans la littérature biostatistique (par exemple, Diggle, Heagerty, Liang et Zeger 2002), la littérature sur la modélisation multinationaux (par exemple, Goldstein 2003) et la littérature économique (par exemple, Baltagi 2001). Nous considérons une classe établie d'estimateurs ponctuels de type moindres carrés généralisés modifiés par les poids de sondage. Pour certaines applications de ces méthodes à des données d'enquête, voir Lavange, Koch et Schwartz (2001), ainsi que Lavange, Stearns, Latala, Koch et Shah (1996).

Nous mesurons l'effet d'un plan d'échantillonnage complexe sur l'estimation de la variance par l'« effet d'erreur de spécification », dénoté *meff* (pour *mispécification effect* en anglais) (Skinner 1989a), qui est la variance de l'estimateur ponctuel d'intérêt sous le plan d'échantillonnage réel divisé par l'espérance d'un estimateur spécifié de la variance. Il s'agit d'une mesure du biais relatif de l'estimateur spécifié de la variance. Si celui-ci est sans biais, le *meff* sera égal à un. Ce concept est étroitement lié à celui de l'« effet de plan » ou *deff* de Kish (1965), défini comme étant la variance de l'estimateur ponctuel sous le plan donné divisé par sa variance sous échantillonnage aléatoire simple avec la même taille d'échantillon, notion qui est plus en rapport avec le choix du plan qu'avec celui de l'estimateur de l'erreur-type.

Nous illustrerons nos arguments théoriques grâce à des analyses de données provenant de la British Household Panel Survey (BHPS) sur les attitudes à l'égard des rôles de l'homme et de la femme, où les principales unités d'intérêt analytique sont les femmes prises individuellement et où les grappes correspondent aux secteurs postaux (*postcode sectors* en anglais) utilisés comme unités primaires d'échantillonnage lors de la sélection de l'échantillon de première vague à partir d'un registre d'adresses.

Le cadre, y compris les modèles et les méthodes d'estimation, est décrit à la section 2. Les propriétés théoriques des méthodes d'estimation de la variance sont exposées à la section 3. La section 4 illustre numériquement ces propriétés, à l'aide d'une analyse des données de la BHPS. Enfin, certaines conclusions sont présentées à la section 5.

2. Modèle de régression, données et méthodes d'inférence

Considérons une population finie $U = \{1, \dots, N\}$ de N unités, que nous supposons fixe au cours d'une série

Estimation de la variance dans l'analyse de données d'enquête longitudinale en grappes

Chris Skinner et Marcel de Toledo Vieira¹

Résumé

Nous étudions l'effet de l'échantillonnage en grappes sur les erreurs-types dans l'analyse des données d'enquête longitudinale. Nous considérons une classe de modèles de régression pour données longitudinales d'usages très répandus et la non-prise en compte de la mise en grappes dans l'estimation de l'erreur-type a tendance à augmenter avec le nombre d'enquêtes sociales. La conséquence est qu'en général, il est au moins aussi important de tenir compte de la mise en grappes dans le calcul des erreurs-types que dans celui des analyses transversales. Nous illustrons cet argument théorique à l'aide des résultats empiriques d'une analyse par régression de données longitudinales sur les attitudes à l'égard des rôles de l'homme et de la femme provenant de l'enquête par panel menée auprès des ménages au Royaume-Uni (*British Household Panel Survey*). Nous comparons aussi deux approches d'estimation de la variance dans l'analyse des données d'enquête longitudinale, à savoir une approche par plan de sondage basée sur la linéarisation et une sous-estime si l'on se contente, en vue d'en tenir compte, d'inclure un effet aléatoire additif pour représenter la mise en grappes dans un modèle multivariés.

Mots clés : Mise en grappes; effet de plan; effet d'erreur de spécification; modèle multivariés.

1. Introduction

Il est bien connu qu'il importe de tenir compte de la mise en grappes de l'échantillon lors de l'estimation des erreurs-types dans l'analyse des données d'enquête. Sinon, les estimateurs des erreurs-types risquent d'être gravement biaisés. Dans le présent article, nous étudions l'effet de la mise en grappes dans l'analyse de données d'enquête longitudinale par régression et le comparons à celui observé dans l'analyse transversale correspondante. Kish et Frankel (1974) ont présenté des travaux empiriques montrant que l'effet d'un plan de sondage complexe sur la variance diminue lorsque les statistiques analytiques deviennent plus complexes et l'on pourrait donc conjecturer que l'effet est susceptible de diminuer aussi dans le cas des analyses longitudinales. Nous soutenons qu'en fait, l'effet de la mise en grappes tend parfois à être plus important dans les analyses longitudinales, du moins pour plusieurs types courants d'analyse et certaines conditions pratiques courantes. Une explication intuitive serait que certaines formes courantes d'analyse longitudinale de données individuelles permettent d'« extraire » de l'estimation des coefficients de régression une grande part de la variation temporelle « aléatoire » présentes dans les réponses individuelles. En revanche, il se peut qu'il soit impossible d'extraire autant de variation des effets de la mise en grappes, puisque cette dernière, représentant la géographie par exemple, a souvent tendance à produire des effets plus stables que les mesures

Il existe une abondante littérature sur les méthodes permettant de tenir compte des plans d'échantillonnage complexes dans l'analyse de données d'enquête par la régression. Voir, par exemple, Kish et Frankel (1974), Fuller (1975), Binder (1983), Skinner, Holt et Smith (1989), ainsi que Chambers et Skinner (2003). Nous ne nous intéressons ici qu'aux analyses par la régression « agrégée » (Skinner et coll. 1989), où les coefficients de régression au « niveau de la population » sont les paramètres d'intérêt, où les estimations appropriées de ces coefficients peuvent être obtenues en adaptant des méthodes basées sur un modèle standard avec l'utilisation de poids de sondage et où les variances de ces coefficients de régression estimés peuvent être estimées par des méthodes de linéarisation (Kish et Frankel 1974; Fuller 1975). Dans le présent article, nous

étudions l'effet de l'échantillonnage en grappes sur la variance de l'estimation de la mise en grappes. Nous étudions la question de la variance de l'estimation de la variance, nous étudions la question de savoir comment entreprendre l'estimation proprement dite de la variance. Il est naturel pour de nombreux analystes de représenter la mise en grappes à l'aide de modèles multivariés. Nous comparons les méthodes d'estimation de la variance fondées sur ce genre de modèle à celles fondées sur le plan de sondage dans le cas de l'échantillonnage en grappes.

Dans son article, Saigo propose une méthode bootstrap d'estimation de la variance pour les plans de sondage à deux phases avec fractions de sondage élevées. La méthode s'appuie sur les techniques du bootstrap courantes, mais comporte un ajustement des valeurs des variables auxiliaires pour les unités qui sont sélectionnées à la première phase seulement. La méthode proposée est illustrée à l'aide de plusieurs estimateurs utilisés couramment, comme l'estimateur par le ratio et les estimateurs de la fonction de répartition et des quantiles. Les résultats d'une étude par simulation comparant la méthode proposée à plusieurs autres sont présentés.

L'article de Longford traite du problème de l'estimation de l'EQM pour les estimations pour petits domaines. L'auteur obtient un estimateur composite de l'EQM des moyennes de petits domaines en combinant un estimateur de la variance sous un modèle et un estimateur naïf de l'EQM. Le coefficient qui combine les deux estimateurs minimise l'EQM prévue de l'estimateur composite de l'EQM résultant. L'estimateur proposé est comparé aux estimateurs existants dans plusieurs études par simulation.

Shao considère le problème de l'imputation pour remplacer les valeurs manquantes en cas de non-réponse non ignorable. Dans la situation où la non-réponse dépend d'un effet aléatoire au niveau de la grappe, il montre que l'estimateur imputé par la moyenne est biaisé, à moins que l'on utilise la moyenne de la grappe. Pour l'estimation de la variance, il fournit une méthode d'estimation de la variance par le jackknife pour l'estimateur proposé. Il compare ce dernier à l'estimateur imputé par la moyenne à l'aide d'une étude par simulation.

Dans le dernier article du numéro, Tiwari, Nigam et Pant utilisent le concept de plan d'échantillonnage proportionnel à la taille le plus proche pour obtenir un plan d'échantillonnage contrôlé optimal assurant que les probabilités de sélection des échantillons non privilégiés soient nulles. Le plan d'échantillonnage contrôlé optimal est obtenu en combinant un plan d'échantillonnage avec probabilité d'inclusion proportionnelle à la taille et des techniques de programmation quadratique pour assurer que les échantillons non privilégiés aient une probabilité de sélection nulle. Les auteurs illustrent leur méthode à l'aide de plusieurs exemples.

Harold Mantel, Rédacteur en chef délégué

Dans ce numéro

Ce numéro de *Techniques d'enquête* comprend des articles portant sur divers sujets méthodologiques tels que la modélisation et l'estimation, la pondération, la non-réponse et l'échantillonnage.

Dans le premier article, Skinner et Vieira étudient l'effet de l'échantillonnage en grappes sur l'estimation de la variance dans les enquêtes longitudinales. Ils présentent des arguments théoriques et des données empiriques démontrant les effets de la non-prise en compte de la mise en grappes dans les analyses longitudinales et constatent qu'en général, ces effets ont tendance à être plus importants que dans le cas des analyses transversales correspondantes. Ils comparent aussi les méthodes basées sur le plan de sondage classiques pour tenir compte de la mise en grappes dans l'estimation de la variance à une approche de modélisation multiniveaux.

Kovacevic et Roberts comparent trois modèles conçus pour l'analyse des périodes multiples émanant de données recueillies au moyen d'enquêtes longitudinales à plan de sondage complexe pouvant comprendre une stratification et une mise en grappes. Ces modèles sont des variantes du modèle à risques proportionnels de Cox du même genre que celles proposées dans la littérature par Lin et Wei (1989), Binder (1992) et Lin (2000). Ces trois modèles sont comparés à l'aide de données provenant de l'Enquête sur la dynamique du travail et du revenu (EDTR) réalisée par Statistique Canada. L'article fournit de nouveaux éclaircissements concernant l'ajustement des modèles de Cox à des données d'enquête représentant plusieurs périodes par individu, situation qui survient assez fréquemment. L'article illustre aussi certains défis que pose l'ajustement des modèles de Cox aux données d'enquête.

Elliott présente dans son article un moyen de réaliser un compromis entre la variance élevée due à des poids de valeur extrême et le biais éventuel à l'aide d'une méthode bayésienne de réduction des poids dans des modèles linéaires généralisés. Le compromis est obtenu en utilisant un modèle hiérarchique bayésien stratifié dans lequel les strates sont déterminées par les probabilités d'inclusion ou par les poids de sondage. Il illustre et évalue l'approche à l'aide de simulations fondées sur des modèles de régression linéaire et de régression logistique, ainsi qu'une application portant sur des données provenant de la Partners for Child Passenger Surveillance Survey.

L'article de Breidt, Opsomer, Johnson et Kanali explore l'utilisation de méthodes semiparamétriques pour l'estimation des moyennes de population. Dans l'estimation semiparamétrique, il est supposé que certaines variables sont reliées linéairement à la variable d'intérêt, tandis que d'autres peuvent l'être de façon plus compliquée, non spécifiée. Les auteurs étudient théoriquement les propriétés sous le plan de sondage des estimateurs résultants. En particulier, ils montrent la convergence sous le plan et la normalité asymptotique de leur estimateur. Puis, ils appliquent leur méthode à des données provenant d'une enquête sur les lacs du Nord-Est des États-Unis.

Tanguay et Lavallée abordent la question de l'estimation de la dépréciation des actifs à l'aide d'une base de données sur les ratios de prix. Dans leur article, le problème est dû au fait que les ratios ne proviennent pas d'un échantillon aléatoire tiré de la population de ratios. Les auteurs soutiennent que la distribution des ratios devrait converger vers une loi uniforme et proposent un scénario de pondération qui rendra la fonction de répartition empirique pondérée approximativement uniforme. Ils illustrent la méthode proposée à l'aide de données sur la dépréciation des automobiles.

Steel et Clark présentent une comparaison empirique et théorique des pondérations produites par la régression généralisée au niveau de la personne et des pondérations intégrées au niveau du ménage dans le cas d'un échantillon aléatoire simple de ménages à partir duquel tous les membres de chaque ménage sont sélectionnés. Ils concluent que l'utilisation de la pondération intégrée est associée à une perte faible, voire nulle, d'efficacité.

∞
The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.

∞
Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 33, numéro 1, juin 2007

Table des matières

Dans ce numéro.....	1
Articles Réguliers	
Chris Skinner et Marcel de Toledo Vieira	
Estimation de la variance dans l'analyse de données d'enquête longitudinale en grappes	3
Milorad S. Kovacević et Georgia Roberts	
Modélisation des durées de périodes multiples à partir de données d'enquête longitudinale.....	15
Michael R. Elliott	
Réduction bayésienne des poids pour les modèles de régression linéaire généralisée.....	27
F. Jay Breidt, Jean D. Opsomer, Alicia A. Johnson et M. Giovanna Ranalli	
Estimation assistée par un modèle semi-paramétrique pour les enquêtes sur les ressources naturelles.....	41
Marc Tanguay et Pierre Lavalée	
Pondération <i>ex post</i> des données de prix pour l'estimation des taux de dépréciation	53
David G. Steel et Robert G. Clark	
Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes-ménages.....	59
Hiroshi Saigo	
Bootstrap avec moyenne ajustée pour l'échantillonnage à deux phases	71
Nicholas Tibor Longford	
De l'erreur-type des estimateurs pour petits domaines fondés sur un modèle.....	81
Jun Shao	
Traitement de la non-réponse dans les sondages en grappes.....	93
Neeraj Tiwari, Arun Kumar Nigam et Ila Pant	
Plan d'échantillonnage proportionnel à la taille le plus proche contrôlé optimal	99

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président D. Royce

Anciens présidents G.J. Brackstone
R. Platek

J. Garbino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

COMITÉ DE RÉDACTION

Rédacteur en chef J. Kovar, *Statistique Canada*
Rédacteur en chef délégué H. Mantel, *Statistique Canada*

Rédacteurs associés

D.A. Binder, *Statistique Canada*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Garbino, *Statistique Canada*
M.A. Hidiroglou, *Statistique Canada*
D. Judkins, *Westat Inc.*
P. Kott, *National Agricultural Statistics Service*
R. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistique Canada*
G. Nathian, *Hebrew University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*
T.J. Rao, *Indian Statistical Institute*
J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importation particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

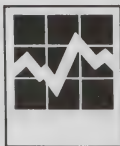
Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en 150 Promenade Timney's Pasture, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ × 2 exemplaires); autres pays, 20 \$ CA (10 \$ × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site internet de Statistique Canada : www.statcan.ca.

Techniques d'enquête



Une revue éditée par Statistique Canada

Juin 2007 • Volume 35 • Numéro 1

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2007

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Juin 2007

N° 12-001-XPB au catalogue

Périodicité : semestrielle

ISSN 0714-0045

Ottawa



Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web à www.statcan.ca.

Service national de renseignements
1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants
1-800-363-7629
Renseignements concernant le Programme des services de dépôt
1-800-889-9734
Renseignements par courriel
infostats@statcan.ca
Site Web
www.statcan.ca

Renseignements pour accéder ou commander le produit

Le produit n° 12-001-XIF au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Publications.

Ce produit n° 12-001-XPB au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

Exemplaire		Abonnement annuel
Etats-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

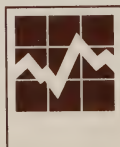
La version imprimée peut être commandée par

- Téléphone (Canada et Etats-Unis) 1-800-267-6677
- Télécopieur (Canada et Etats-Unis) 1-877-287-4369
- Courriel infostats@statcan.ca
- Poste Statistique Canada
- Division des finances
- Immeuble R.-H.-Coats, 6^e étage
- 100, promenade Tunney's Pasture
- Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de nous > Offrir des services aux Canadiens.



Techniques d'enquête

Une revue
éditée
par Statistique Canada

N° 12-001-XPB au catalogue

Juin 2007

Volume 33

Numéro 1



Statistique
Canada
Statistics
Canada

Canada

12-001



Government
Publications

Survey Methodology

Catalogue No. 12-001-XPB

A journal
published by
Statistics Canada

December 2007

•

Volume 33

•

Number 2



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.ca or contact us by e-mail at infostats@statcan.ca or by telephone from 8:30 a.m. to 4:30 p.m. Monday to Friday:

Statistics Canada National Contact Centre

Toll-free telephone (Canada and the United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369
Depository Services Program inquiries line	1-800-635-7943
Depository Services Program fax line	1-800-565-7757

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Accessing and ordering information

This product, Catalogue no. 12-001-X, is available for free in electronic format. To obtain a single issue, visit our website at www.statcan.ca and select "Publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered by:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.ca
- Mail
Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

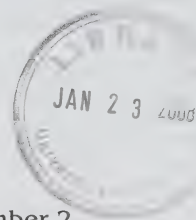
Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1-800-263-1136. The service standards are also published on www.statcan.ca under "About us" > "Providing services to Canadians."



Survey Methodology

A journal
published by
Statistics Canada



December 2007 • Volume 33 • Number 2

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2007

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows: Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 2007

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman D. Royce

Past Chairmen G.J. Brackstone
R. Platek

Members J. Gambino
R. Jones
J. Kovar
H. Mantel
E. Rancourt

EDITORIAL BOARD

Editor J. Kovar, *Statistics Canada*
Deputy Editor H. Mantel, *Statistics Canada*

Past Editor M.P. Singh

Associate Editors

D.A. Binder, *Statistics Canada*
J.M. Brick, *Westat Inc.*
P. Cantwell, *U.S. Bureau of the Census*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
M.A. Hidirolou, *Statistics Canada*
D. Judkins, *Westat Inc.*
P. Kott, *National Agricultural Statistics Service*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
G. Nathan, *Hebrew University*
J. Opsomer, *Colorado State University*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *University of Southampton*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
V.J. Verma, *Università degli Studi di Siena*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
A. Zaslavsky, *Harvard University*

Assistant Editors J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.ca.

Survey Methodology
A Journal Published by Statistics Canada
Volume 33, Number 2, December 2007

Contents

In This Issue.....	95
 Waksberg Invited Paper Series	
Carl-Erik Särndal The calibration approach in survey theory and practice	99
 Regular Papers	
Seppo Laaksonen Weighting for two-phase surveyed data	121
Pascal Ardilly and Pierre Lavallée Weighting in rotating samples: The SILC survey in France	131
Jay J. Kim, Jianzhu Li and Richard Valliant Cell collapsing in poststratification	139
Fulvia Mecatti A single frame multiplicity estimator for multiple frame surveys	151
David Haziza Variance estimation for a ratio in the presence of imputed data	159
James Chipperfield and John Preston Efficient bootstrap for business surveys	167
Jacob J. Oleson, Chong Z. He, Dongchu Sun and Steven L. Sheriff Bayesian estimation in small areas when the sampling design strata differ from the study domains	173
Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacei Small area estimation of average household income based on unit level models for panel data.....	187
Anne Renaud Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results	199
Marc N. Elliott and Amelia Haviland Use of a web-based convenience sample to supplement a probability sample	211
Acknowledgements.....	217

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.



In This Issue

This issue of *Survey Methodology* opens with the seventh paper in the annual invited paper series in honour of Joseph Waksberg. The editorial board would like to thank the members of the selection committee - Gordon Brackstone, chair, Bob Groves, Sharon Lohr and Wayne Fuller – for having selected Carl-Erik Särndal as the author of this year's Waksberg Award paper. For this occasion, a special Workshop on Calibration and Estimation in Surveys (WCES) was organised on October 31st and November 1st at Statistics Canada. Professor Carl-Erik Särndal was the keynote speaker and presented his Waksberg paper. During the two days, 12 other speakers presented a paper and paid their tribute to Carl-Erik Särndal.

In his paper entitled "The Calibration Approach in Survey Theory and Practice" Särndal discusses the development and application of calibration in survey sampling. He describes the concept of calibration in some detail and contrasts it with generalized regression. He then describes different approaches to calibration including the minimum distance method, instrumental variables, and model calibration. Several examples of calibration and alternatives are considered.

Laaksonen discusses weighting in two phase sampling in which respondents to the first phase are asked if they are willing to participate in the second phase. The weighting thus has to deal with non-response at both phases of the survey, and also account for first-phase respondents who were unwilling to participate in the second phase. Using data from a Finnish survey on leisure-time activities, he empirically evaluates variations on a weighting method that uses response propensity modeling and calibration.

The article by Ardilly and Lavallée discusses the weighting problem for the SILC (Statistics on Income and Living Conditions) survey in France. This survey uses a rotating sample plan with nine panels. To obtain approximate estimators without bias, the authors relied on the weight-share method. Longitudinal weighting is discussed first, and then cross-sectional weighting is also discussed.

The paper by Kim, Li and Valliant deals with the problem of small cells or large weight adjustments when poststratification is used. The authors first describe several standard estimators and then introduce two alternative estimators based on cell collapsing. They study the performance of these estimators in terms of their effectiveness in controlling the coverage bias and the design variance. These properties are evaluated theoretically and also through a simulation study using a population based on the 2003 National Health Interview Survey.

Mecatti proposes a simple multiplicity estimator in the context of multi-frame surveys. She first shows that the proposed estimator is design-unbiased. Then, she proposes an unbiased estimator of the variance of the multiplicity estimator. Using 29 simulated populations, she compares the multiplicity estimator with alternative estimators proposed in the literature.

Haziza studies the problem of variance estimation for a ratio of two totals when marginal random hot deck imputation has been used to fill in missing data. Two approaches to inference are considered, one using an imputation model and a second one using a nonresponse model. Variance estimators are derived under two frameworks: the reverse approach of Shao and Steel (1999) and the traditional two-phase approach.

In their paper, Chipperfield and Preston describe the without replacement scaled bootstrap variance estimator that was implemented in the Australian Bureau of Statistics' generalized estimation system ABSEST. The without replacement scaled bootstrap estimator is shown to be more efficient than the with replacement scaled bootstrap estimator for stratified samples when the stratum sizes are small. In addition, the without replacement scaled bootstrap estimator was shown to require fewer replicates to achieve the same replication error as the with replacement estimator. For the ABSEST system, bootstrap variance estimators were chosen over other variance estimation methods for their computational efficiency and the without replacement bootstrap was selected for the reasons above.

Oleson, He and Sun describe a Bayesian modelling approach for situations where the sampling design is stratified and the estimation procedure requires post-stratification. The method is illustrated with data from the 1998 Missouri Turkey Hunting Survey for which the strata were defined by the hunter's place of residence but estimates were required at the county level.

Fabrizi, Ferrante and Pacei discuss a methodology which is increasingly important in modern sample survey applications. They investigate the effect of borrowing strength from additional panel information for cross sectional household income estimates for small areas in Italy. The proposed methods seem to tackle a problem which may have further relevance for European Official Statistics, and possibly also in the area of small area statistics for indicators which may be used for policy research.

Renaud presents an interesting application of a post-enumeration survey to estimate net undercoverage in the 2000 census in Switzerland. The objective of this survey was slightly different from that of other countries in that it was not designed to adjust the Census counts for net undercoverage, but rather to gather information to improve the quality of subsequent censuses.

In the final paper, Elliot and Haviland consider combining a convenience sample with a probability based sample to obtain an estimate with a smaller MSE. The resulting estimator is a linear combination of the convenience and probability sample estimates with weights that are a function of the bias. By looking at the maximum incremental contribution of the convenience sample, they show that improvement to the MSE may be attainable only in certain circumstances.

Harold Mantel, Deputy Editor

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work. The author receives a cash award made possible by a grant from Westat, in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially by the American Statistical Association. Previous winners are listed below. Their papers in the series have already appeared in *Survey Methodology*.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)

Nominations:

The author of the 2009 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, Robert Groves, by email to bgroves@isr.umich.edu. Nominations and suggestions for topics must be received by February 29, 2008.

2007 Waksberg Invited Paper

Author: Carl-Erik Särndal

Carl-Erik Särndal, retired professor in the Université de Montréal, is a consultant and expert who has been associated with several national statistical institutes, in particular Statistics Canada and Statistics Sweden as well as Statistics Finland, INSEE and Eurostat. His list of publications comprises three books, including the very well reknown Model Assisted Survey Sampling book that has had a major impact. He is also the author of numerous scientific articles, in sole authorship or in collaboration with researchers from many countries. His research interest in survey sampling has been very diversified, but most often revolved around ways to best using auxiliary information in sampling and estimation.

Members of the Waskberg Paper Selection Committee (2007-2008)

Robert Groves, (Chair)

Wayne A. Fuller, *Iowa State University*

Daniel Kasprzyk, *Mathematica Policy Research*

Leyla Mojadjer, *Westat*

Past Chairs:

Graham Kalton (1999 - 2001)

Chris Skinner (2001 - 2002)

David A. Binder (2002 - 2003)

J. Michael Brick (2003 - 2004)

David R. Bellhouse (2004 - 2005)

Gordon Brackstone (2005 - 2006)

Sharon Lohr (2006 - 2007)

The calibration approach in survey theory and practice

Carl-Erik Särndal¹

Abstract

Calibration is the principal theme in many recent articles on estimation in survey sampling. Words such as “calibration approach” and “calibration estimators” are frequently used. As article authors like to point out, calibration provides a systematic way to incorporate auxiliary information in the procedure.

Calibration has established itself as an important methodological instrument in large-scale production of statistics. Several national statistical agencies have developed software designed to compute weights, usually calibrated to auxiliary information available in administrative registers and other accurate sources.

This paper presents a review of the calibration approach, with an emphasis on progress achieved in the past decade or so. The literature on calibration is growing rapidly; selected issues are discussed in this paper.

The paper starts with a definition of the calibration approach. Its important features are reviewed. The calibration approach is contrasted with (generalized) regression estimation, which is an alternative but conceptually different way to take auxiliary information into account. The computational aspects of calibration are discussed, including methods for avoiding extreme weights. In the early sections of the paper, simple applications of calibration are examined: The estimation of a population total in direct, single phase sampling. Generalization to more complex parameters and more complex sampling designs are then considered. A common feature of more complex designs (sampling in two or more phases or stages) is that the available auxiliary information may consist of several components or layers. The uses of calibration in such cases of composite information are reviewed. Later in the paper, examples are given to illustrate how the results of the calibration thinking may contrast with answers given by earlier established approaches. Finally, applications of calibration in the presence of nonsampling error are discussed, in particular methods for nonresponse bias adjustment.

Key Words: Auxiliary information; Weighting; Consistency; Design-based inference; Regression estimator; Models; Nonresponse; Complex sampling design.

1. Introduction

1.1 Calibration defined

It is useful in this paper to refer to a definition of the calibration approach. I propose the following formulation.

Definition. The calibration approach to estimation for finite populations consists of

- (a) a computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),
- (b) the use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units,
- (c) an objective to obtain nearly design unbiased estimates as long as nonresponse and other non-sampling errors are absent.

In the literature, “calibration” frequently refers to (a) alone; I shall often use the term for (a) to (c) together. Earlier definitions, although less extensive, agree essentially with mine. Ardilly (2006) defines calibration (or, more precisely, “calage généralisé”) as a method of re-weighting used when one has access to several variables, qualitative or

quantitative, on which one wishes to carry out, jointly, an adjustment.

Kott (2006) defines calibration weights as a set of weights, for units in the sample, that satisfy a calibration to known population totals, and such that the resulting estimator is randomization consistent (design consistent), or, more rigorously, that the design bias is, under mild conditions, an asymptotically insignificant contribution to the estimator’s mean squared error. This is the property I call “nearly design unbiased”.

The Quality Guidelines (fourth edition) of Statistics Canada (2003) say: “Calibration is a procedure than can be used to incorporate auxiliary data. This procedure adjusts the sampling weights by multipliers known as calibration factors that make the estimates agree with known totals. The resulting weights are called calibration weights or final estimation weights. These calibration weights will generally result in estimates that are design consistent, and that have a smaller variance than the Horvitz-Thompson estimator.”

Part (c) of the definition merits a comment. Nothing prevents producing weights calibrated to given auxiliary information without requiring (c). But most published work on calibration is in the spirit of (c), so it makes good sense to include it. When non-sampling errors are present, bias in the estimates is unavoidable, whether they are made by calibration or by any other method. In line with (c), I

1. Carl-Erik Särndal, 2115 Erinbrook Crescent, #44, Ottawa, Ontario, K1B 4J5, Canada. E-mail: carl.sarndal@rogers.com.

consider design-based inference to be the standard in this paper. The randomization-based variance of an estimator is thus important. However, the paper focuses on “motivations behind (point) estimation”; for reasons of space, the important question of variance estimation is not addressed.

1.2 Comments arising

The definition in Section 1.1 prompts some comments and references to earlier literature:

(1) *Calibration as a linear weighting method.* Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies.

Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horvitz and Thompson (1952). Weighting became widely accepted. Later, post-stratification weighting achieved the same status. Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent; the weights depend on the observed sample.

Inverse inclusion probability weights are, by definition, greater than or equal to unity. A commonly heard interpretation is that “an observed unit represents itself and a number of others, not observed”. Calibrated weights, on the other hand, are not necessarily greater than or equal to unity, unless special care is taken in the computation to obtain this property.

Calibration is new as a term in survey sampling - about 15 years old - but not as a technique for producing weights. Those who maintain “I practiced calibration long before it was called calibration” have a point. The last 15 years widened the scope and the appeal of the technique. Weighting akin to calibration has long been used by private survey institutes, for example, in connection with quota sampling, a form of non-probability sampling outside the scope of this paper.

Weighting of observed variable values was an important topic before calibration became a popular term. Some authors derived the weights via the argument that they should differ as little as possible from the unbiased sampling design weights (the inverse of the inclusion probabilities). Others found the weights by recognizing that a linear regression estimator can be written as a linearly weighted sum of the observed study variable values. Terms such as

“survey sample weighting” and “regression weighting” and “case weighting” are used. Among such “early papers” are Alexander (1987), Bankier, Rathwell and Majkowski (1992), Bethlehem and Keller (1987), Chambers (1996), Fuller, Loughin and Baker (1994), Kalton and Flores-Cervantes (1998), Lemaître and Dufour (1987), Särndal (1982) and Zieschang (1990). I comment later on the technique “repeated weighting”, promoted by the Dutch national statistical agency, CBS. The newer term “calibration” conveys a more specific message and a more definite direction than the older “weighting”.

(2) *Calibration as a systematic way to use auxiliary information.* Calibration provides a systematic way to take auxiliary information into account. As Rueda, Martínez, Martínez and Arcos (2007) point out, “in many standard settings, the calibration provides a simple and practical approach to incorporating auxiliary information into the estimation”.

Auxiliary information was used to improve the accuracy of survey estimates long before calibration became popular. Numerous papers were written with this goal in mind, for more or less specialized situations. Today, calibration does offer a systematic outlook on the uses of auxiliary information. For example, calibration can deal effectively with surveys where auxiliary information exists at different levels. In two-stage sampling information may exist for the first stage sampling units (the clusters), and other information for the second stage sampling units. In surveys with nonresponse (that is, essentially all surveys), information may exist “at the population level” (known population totals), and other information “at the sample level” (auxiliary variable values for all those sampled, responding and non-responding). Calibration with “composite information” is reviewed in Sections 8 and 9.

Regression estimation, or generalized regression (GREG) estimation, competes with calibration as a systematic way to incorporating auxiliary information. It is therefore important to contrast GREG estimation (described in Section 3) with calibration estimation (described in Section 4). The two approaches are different.

(3) *Calibration to achieve consistency.* Calibration is often described as “a way to get consistent estimates”. (Here “consistent” refers not to “randomization consistent” but to “consistent with known aggregates”.) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. A desire to promote credibility in published statistics is an often cited reason for demanding consistency. Some users of statistics dislike finding the same population

quantity estimated by two or more numbers that do not agree.

The totals with which consistency is sought are sometimes called control totals. “Controlled weights” or “calibrated weights” suggest improved, more accurate estimation. The French term for calibration, “calage”, has a similar connotation of “stability”.

Consistency through calibration has a broader implication than just agreement with known population auxiliary totals. Consistency can, for example, be sought with appropriately estimated totals, arising in the current survey or in other surveys.

Consistency among tables estimated from different surveys is the motive behind *repeated weighting*, the technique developed at the Dutch national statistical agency CBS in several articles: Renssen and Nieuwenbroek (1997); Nieuwenbroek, Renssen, and Hofman (2000); Renssen, Kroese, and Willeboordse (2001); Kottnerus and van Duin (2006). The stated objective is to accommodate user demands to produce numerically consistent outputs. As the last mentioned paper points out, repeated weighting can be seen as an additional calibration step for a new adjustment of already calibrated weights. The final weights realize consistency with given margins.

Consistency with known or estimated totals may bring the extra benefit of improved accuracy (lower variance and/or reduced nonresponse bias). However, in some articles, especially those authored in statistical agencies, consistency for user satisfaction seems a more imperative motivation than the prospect of increased accuracy.

When the primary motivation for calibration is not so much an agreement with other statistics as rather to reduce variance and/or nonresponse bias, then “balanced weight system” is a more appropriate description than “consistent weight system”, because the objective is then to balance the weights to reflect the outcome of the sampling, the response to the survey, and the information available.

(4) *Calibration for convenience and transparency.* As Harms and Duchesne (2006) point out, “The calibration approach has gained popularity in real applications because the resulting estimates are easy to interpret and to motivate, relying, as they do, on design weights and natural calibration constraints.” Calibration on known totals strikes the typical user as transparent and natural. Users who understand sample weighting appreciate that calibration leaves the design weights “slightly modified only”, while respecting the controls. The unbiasedness is only negligibly disturbed. The simpler forms of calibration invoke no assumptions, only “natural constraints”. Yet another advantage is appreciated by users: In many applications, calibration gives a unique weighting system, applicable to

all study variables, of which there are usually many in large government surveys.

(5) *Calibration in combination with other terms.* Some authors use the word “calibration” in combination with other terms, to describe various directions of thought. Examples of this proliferation of terms are: Model-calibration (Wu and Sitter 2001); g-calibration (Vanderhoeft, Waeytens and Museux 2000); Harmonized calibration (Webber, Latouche and Rancourt 2000); Higher level calibration (Singh, Horn and Yu 1998); Regression calibration (Demnati and Rao 2004); Non-linear calibration (Plikusas 2006); Super generalized calibration (Calage super généralisé; Ardilly 2006); Neural network model-calibration estimator and Local polynomial model-calibration estimator (Montanari and Ranalli 2003, 2005); Model-calibrated pseudo empirical maximum likelihood estimator (Wu 2003), and yet others. Also, calibration plays a significant role in the indirect sampling methods proposed in Lavalée (2006). In a somewhat different spirit, not reviewed here, are concepts such as calibrated imputation (Beaumont 2005a), and bias calibration (Chambers, Dorfman and Wehrly (1993), Zheng and Little (2003)). The following review pages do not give justice to all the innovations within the sphere of calibration, but the names alone do suggest directions that have been explored.

(6) *Calibration as a new direction for thought.* If calibration represents “a new approach” with clear differences compared with predecessors, we must examine such questions as: Does calibration generalize earlier theories or approaches? Does calibration give better, more satisfactory answers on questions of importance, as compared with earlier recognized approaches? Sections 4.5 and 7.1 in this paper illustrate how the answers provided by calibration compare with, or contrast with, those obtained in earlier modes of reasoning.

The practice of survey sampling encounters “nuisances” such as nonresponse, frame deficiencies and measurement errors. It is true that imputation and reweighting for nonresponse are widely practiced, through a host of techniques. But they are somehow “separate issues”, still waiting to be more fully embedded into a comprehensive, more satisfactory theory of inference in sample surveys. Many theory papers deal with estimation for an imagined ideal survey, nonexistent in practice, where nonresponse and other non-sampling errors are absent. This is not a criticism of the many excellent but idealized theory papers. The foundations need to be explored, too.

Sections 9 and 10 indicate that calibration can provide a more systematic outlook on inference in surveys even in the presence of the various non-sampling errors. Future fruitful developments are expected in that regard.

2. Basic conditions for design-based estimation in sample surveys

This section sets the background for Sections 3 to 7. By “basic conditions” I will mean single phase probability sampling of elements and full response. In practice, survey conditions are not that simple and perfect, but many theory papers nevertheless address this situation.

A probability sample s is drawn from the finite population $U = \{1, 2, \dots, k, \dots, N\}$. The probability sampling design generates for element k a known inclusion probability, $\pi_k > 0$, and a corresponding sampling design weight $d_k = 1/\pi_k$. The value y_k of the study variable y is recorded for all $k \in s$ (complete response). The objective is to estimate a population total $Y = \sum_U y_k$ with the use of auxiliary information. The study variable y may be continuous or, as in many government surveys, categorical. For example, if y is dichotomous with value $y_k = 0$ or $y_k = 1$ according as person k is employed or unemployed, then the parameter $Y = \sum_U y_k$ to be estimated is the population count of unemployed people. (If $A \subseteq U$ is a set of elements, I write \sum_A for $\sum_{k \in A}$.) The basic design unbiased estimator of Y is $\hat{Y}_{HT} = \sum_s d_k y_k$, the Horvitz-Thompson estimator. It is, however, inefficient when powerful auxiliary information is available for use at the estimation phase.

The general notation for the auxiliary vector will be \mathbf{x}_k . In some countries, for some surveys, the sources of auxiliary data permit extensive vectors \mathbf{x}_k to be built. But some examples of simple vectors are: (1) $\mathbf{x}_k = (1, x_k)'$, where x_k is the value for element k of a continuous auxiliary variable x ; (2) the classification vector used to code membership in one of P mutually exclusive and exhaustive groups, $\mathbf{x}_k = \gamma_k = (\gamma_{1k}, \dots, \gamma_{pk}, \dots, \gamma_{Pk})'$, so that, for $p = 1, 2, \dots, P$, $\gamma_{pk} = 1$ if k belongs to group p , and $\gamma_{pk} = 0$ if not; (3) the combination of (1) and (2), $\mathbf{x}_k = (\gamma_k', x_k \gamma_k')'$; (4) the vector \mathbf{x}_k that codifies two classifications strung out ‘side-by-side’, the dimension of \mathbf{x}_k being $P + Q - 1$, where P and Q are the respective number of categories, and the ‘minus-one’ is to avoid a singular matrix in the computation of weights calibrated “to the margins”; (5) the extension of (4) to more than two ‘side-by-side’ categorical classifications. Cases 4 and 5 are particularly important for production in national statistical agencies.

In auxiliary reasoning it is crucially important to specify exactly the *auxiliary information*. Under the basic conditions we need to distinguish two different cases relative to \mathbf{x}_k :

- (i) \mathbf{x}_k is a known vector value for every $k \in U$ (complete auxiliary information)
- (ii) $\sum_U \mathbf{x}_k$ is known (imported) total, and \mathbf{x}_k is known (observed) for every $k \in s$

It is often the survey environment that dictates whether (i) or (ii) prevails. Case (i), complete auxiliary information, occurs when \mathbf{x}_k is specified in the sampling frame for every $k \in U$ (and thus known for every $k \in s$). This environment is typical of surveys on individuals and households in Scandinavia and other North European countries equipped with high quality administrative registers that can be matched with the frame to provide a large number of potential auxiliary variables. The population total $\sum_U \mathbf{x}_k$ is obtained simply by adding the \mathbf{x}_k .

Case (i) gives considerable freedom in structuring the auxiliary vector \mathbf{x}_k . For example, if x_k is a continuous variable value specified for every $k \in U$, then we are invited to consider x_k^2 and other functions of x_k for inclusion in \mathbf{x}_k , because totals such as $\sum_U x_k^2$ and $\sum_U \log x_k$ are readily computed. If the relationship to the study variable y is curved, it may be a serious omission not to take into account known totals such as the quadratic one or the logarithmic one.

Case (ii) prevails in surveys where (i) is not met, but where $\sum_U \mathbf{x}_k$ is imported from an outside source considered accurate enough, and the individual value \mathbf{x}_k is available (observed in data collection) for every $k \in s$. Then $\sum_U \mathbf{x}_k$ is sometimes called an “independent control total”, to mark its origin from outside the survey itself. Case (ii) is less flexible: If x_k is a variable with a total $\sum_U x_k$ imported from a reliable source, then $\sum_U x_k^2$ may be unavailable, barring x_k^2 from inclusion into \mathbf{x}_k .

3. Generalized regression estimation under the basic conditions

3.1 The GREG concept

Before examining calibration, let us consider *generalized regression* (GREG) *estimation* (or just *regression estimation*), for two good reasons: (1) GREG estimation can also be claimed to be a systematic way to take auxiliary information into account; (2) some (but not all) GREG estimators are calibration estimators, in that they can be expressed in terms of a calibrated linear weighting.

GREG estimators and calibration estimators have been extensively studied in the last two decades. The terms alone, “GREG estimation” and “calibration estimation”, reflect a clear difference in thinking. Statisticians who work in the area are of two types: Those dedicated to “GREG thinking” and those dedicated to “calibration thinking”. The distinction may not be completely clear-cut, but it helps structuring this review paper, so I will use it. I am not venturing to say that the latter thinking is more prevalent in national statistical agencies and the former more prevalent in the academic circles, but perhaps there is such a tendency.

The GREG estimator concept evolved gradually since the mid-1970's. The simple (linear) GREG is explained in Särndal, Swensson and Wretman (1992); a thorough review of regression estimation is given in Fuller (2002). The central idea is that predicted y -values \hat{y}_k can be produced for all N population elements, via the fit of an *assisting model* and the use of the auxiliary vector values \mathbf{x}_k known for all $k \in U$. The predicted values serve to build a nearly design unbiased estimator of the population total $Y = \sum_U y_k$ as

$$\begin{aligned}\hat{Y}_{\text{GREG}} &= \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k) \\ &= \sum_s d_k y_k + \left(\sum_U \hat{y}_k - \sum_s d_k \hat{y}_k \right).\end{aligned}\quad (3.1)$$

The obvious motivation behind this construction is the prospect of a highly accurate estimate \hat{Y}_{GREG} through a close fitting assisting model that leaves small residuals $y_k - \hat{y}_k$. That modeling is the corner stone of GREG thinking. Some authors use the (also justifiable) name *general difference estimator* for the construction (3.1).

The great variety of possible assisting models generates a wide family of GREG estimators of the form (3.1). The assisting model, an imagined relationship between \mathbf{x} and y , can have many forms: linear, non-linear, generalized linear, mixed (model with some fixed, some random effects), and so on. Whatever the choice, the model is "assisting only"; even though it may be short of "true", (3.1) is nearly design unbiased under mild conditions on the assisting model and on the sampling design, so that $(\hat{Y}_{\text{GREG}} - Y)/N = O_p(n^{-1/2})$ and $(\hat{Y}_{\text{GREG}} - Y)/N = (\hat{Y}_{\text{GREG,lin}} - Y)/N + O_p(n^{-1})$, where the statistic $\hat{Y}_{\text{GREG,lin}}$, the result of linearizing \hat{Y}_{GREG} , is unbiased for Y .

3.2 Linear GREG

By linear GREG I mean one that is generated by a linear fixed effects assisting model. The predictions are $\hat{y}_k = \mathbf{x}_k' \mathbf{B}_{s;dq}$ with

$$\mathbf{B}_{s;dq} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_s d_k q_k \mathbf{x}_k y_k \right)$$

so (3.1) becomes

$$\hat{Y}_{\text{GREG}} = \left(\sum_U \mathbf{x}_k \right)' \mathbf{B}_{s;dq} + \sum_s d_k (y_k - \mathbf{x}_k' \mathbf{B}_{s;dq}). \quad (3.2)$$

The q_k are scale factors, chosen by the statistician. The standard choice is $q_k = 1$ for all k . The choice of the q_k has some (but often limited) impact on the accuracy of \hat{Y}_{GREG} ; near-unbiasedness holds for any specification (barring outrageous choices) for the q_k . Although the model is simple, the linear GREG (3.2) contains many estimators, considering the many possible choices of the auxiliary vector \mathbf{x}_k and the scale factors q_k . Under general conditions,

$$(\hat{Y}_{\text{GREG}} - Y)/N = \left(\sum_s d_k E_k - \sum_U E_k \right) / (N + O_p(n^{-1}))$$

where $\sum_s d_k E_k$ is the Horvitz-Thompson estimator in the residuals $E_k = y_k - \mathbf{x}_k' \mathbf{B}_{U;q}$ with $\mathbf{B}_{U;q} = (\sum_U q_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U q_k \mathbf{x}_k y_k)$. Hence, the design-based properties $E(\hat{Y}_{\text{GREG}}) \approx Y$ and $\text{Var}(\hat{Y}_{\text{GREG}}) \approx \text{Var}(\sum_s d_k E_k)$. A close fitting linear regression of y on \mathbf{x} holds the key to a small variance for \hat{Y}_{GREG} (and this is very different from claiming that "a linear regression is the true regression").

The linear GREG in Särndal, Swensson and Wretman (1992) was motivated via the linear assisting model ξ stating that $E_\xi(y_k) = \mathbf{b}' \mathbf{x}_k$ and $\text{Var}_\xi(y_k) = \sigma_k^2$. Generalized least squares fit gives the estimator (3.2) with $q_k = 1/\sigma_k^2$. In that context, an educated guess about the variation of the residuals $y_k - \mathbf{b}' \mathbf{x}_k$ determines the q_k . When the vector \mathbf{x}_k is fixed, the modeling effort boils down to an opinion about the residual pattern. The choice $\sigma_k^2 = \sigma^2 x_k$ gives the classical ratio estimator. If $q_k = \mathbf{m}' \mathbf{x}_k$ for all $k \in U$ and a constant vector \mathbf{m} , then (3.2) reduces to "the cosmetic form" $(\sum_U \mathbf{x}_k)' \mathbf{B}_{s;dq}$.

As Beaumont and Alavi (2004) and others have pointed out, the linear GREG estimator is bias-robust (nearly unbiased although the assisting model falls short of "correct"), but it can be considerably less efficient (have larger mean squared error) than model dependent alternatives which, although biased, may have a considerably smaller variance. Thus one may claim that linear GREG is not variance robust; nevertheless, it is a basic concept in design-based survey theory.

The specification of \mathbf{x}_k should include variables (with known population totals) that served already in defining the sampling design. Design stage information should not be relinquished at the estimation stage; instead, a "repeated usage" is recommended. For example, in stratified simple (STSI) random sampling, the vector \mathbf{x}_k in estimator (3.2) should include, along with other available variables, the dummy coded stratum identifier, $\gamma_k = (\gamma_{k1}, \gamma_{k2}, \dots, \gamma_{kh}, \dots, \gamma_{kH})'$, where $\gamma_{kh} = 1$ if element k belongs to stratum h , and $\gamma_{kh} = 0$ if not; $h = 1, \dots, H$.

We can write the linear GREG (3.2) as a weighted sample sum, $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, with

$$w_k = d_k g_k; \quad g_k = 1 + q_k \lambda' \mathbf{x}_k;$$

$$\lambda' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)' \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1}. \quad (3.3)$$

The weights w_k happen to be *calibrated* to (consistent with) the known population \mathbf{x} -total: $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. That \hat{Y}_{GREG} is expressible as a linearly weighted sum with calibrated weights is a fortuitous by-product. It is not part of GREG thinking, whose central idea formulated in (3.1) is the fit of an assisting model. A few other GREG's than the

simple linear one also have the calibration property, as will be noted later.

3.3 Non-linear GREG

Two features of the linear GREG (3.2) make it a favourite choice for routine production in statistical agencies: (i) the auxiliary population total $\sum_U \mathbf{x}_k$ becomes factored out, so the estimation can proceed as long as an accurate value for that total can be computed or imported, and (ii) when written as the linearly weighted sum $\hat{Y}_{\text{GREG}} = \sum_s w_k y_k$, the weight system (3.3) is independent of the y -variable and can thereby be applied to all y -variables in the survey. We need not know \mathbf{x}_k individually for all $k \in U$; knowing $\sum_U \mathbf{x}_k$ suffices. Needless to say, if we do know all \mathbf{x}_k , more efficient (still nearly design unbiased) members of the GREG family (3.1) can be sought. This will also counter another criticism of the linear GREG, namely that a linear model is unrealistic for some types of data. For example, for a dichotomous y -variable, a logistic assisting model may be both more realistic and yield a more precise GREG estimator.

By a non-linear GREG estimator I mean one generated as in (3.1) by an assisting model of other type than “linear in \mathbf{x}_k with fixed effects”. Among the first to extend the GREG concept in this direction are Firth and Bennett (1998) and Lehtonen and Veijanen (1998); see also Chambers *et al.* (1993). In the last few years, several authors have studied model-assisted non-linear GREG’s.

Non-linear GREG is a versatile idea; a variety of estimators become possible via assisting models ξ of the following type:

$$E_\xi(y_k | \mathbf{x}_k) = \mu_k \quad \text{for } k \in U \quad (3.4)$$

where the model mean μ_k and the model variance $V_\xi(y_k | \mathbf{x}_k)$ are given appropriate formulations.

One application of (3.4) is when $\mu_k = \mu(\mathbf{x}_k, \boldsymbol{\theta})$ is a specified non-linear function in \mathbf{x}_k . Having estimated $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, the fitted values needed for \hat{Y}_{GREG} in (3.1) are $\hat{y}_k = \mu(\mathbf{x}_k, \hat{\boldsymbol{\theta}})$ for $k \in U$. For example, if the modeler specifies $\log \mu_k = \alpha + \beta x_k$, the predictions for use in (3.1) are, following parameter estimation, $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$.

Other applications of (3.4) include generalized linear models such that $g(\mu_k) = \mathbf{x}_k' \boldsymbol{\theta}$, for a specified link function $g(\cdot)$, and $V_\xi(y_k | \mathbf{x}_k) = v(\mu_k)$ is given an appropriate structure. We estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$, the fitted values needed for the non-linear GREG estimator (3.1) are $\hat{y}_k = \hat{\mu}_k = g^{-1}(\mathbf{x}_k' \hat{\boldsymbol{\theta}})$. For example, using a logistic assisting model, $\mathbf{x}_k' \boldsymbol{\theta} = \text{logit}(\mu_k) = \log(\mu_k / (1 - \mu_k))$, and $\hat{y}_k = \hat{\mu}_k = \exp(\mathbf{x}_k' \hat{\boldsymbol{\theta}}) / (1 + \exp(\mathbf{x}_k' \hat{\boldsymbol{\theta}}))$.

Lehtonen and Veijanen (1998) examine the case of a categorical study variable with I classes, $i = 1, 2, \dots, I$, $y_{ik} = 1$ if element k belongs to category i , and $y_{ik} = 0$ if

not. For example, in a Labour Force Survey with $I = 3$ categories, “employed”, “not employed” and “not in the labour force”, an objective is to estimate the respective population counts $Y_i = \sum_U y_{ik}$, $i = 1, 2, 3$. These authors use the logistic assisting model

$$E_\xi(y_{ik} | \mathbf{x}_k) = \mu_{ik}; \mu_{ik} = \exp(\mathbf{x}_k' \boldsymbol{\theta}_i) / \left(1 + \sum_{i=2}^I \exp(\mathbf{x}_k' \boldsymbol{\theta}_i) \right). \quad (3.5)$$

Estimates $\hat{\boldsymbol{\theta}}_i$ of the $\boldsymbol{\theta}_i$ are obtained by maximizing design-weighted log-likelihood. The resulting predictions $\hat{y}_{ik} = \hat{\mu}_{ik}$ are used to form $\hat{Y}_{i \text{ GREG}} = \sum_U \hat{y}_{ik} + \sum_s d_k(y_{ik} - \hat{y}_{ik})$, for $i = 1, 2, \dots, I$.

Another development is the application of GREG reasoning to estimation for domains, as in Lehtonen, Särndal and Veijanen (2003, 2005) and Myrskylä (2007). Mixed models are used in the first two of these papers to assist the non-linear GREG. Let U_a be a domain, $U_a \subset U$, whose total $Y_{ia} = \sum_{U_a} y_{ik}$ we wish to estimate, $i = 1, 2, \dots, I$. The 2005 paper derives the predictions for the non-linear GREG from the logistic mixed model stating that for $k \in U_a$

$$E_\xi(y_{ik} | \mathbf{x}_k; \mathbf{u}_{ia}) = \exp(\mathbf{x}_k' \boldsymbol{\theta}_{ia}) / \left(1 + \sum_{i=2}^I \exp(\mathbf{x}_k' \boldsymbol{\theta}_{ia}) \right) \quad (3.6)$$

with $\boldsymbol{\theta}_{ia} = \boldsymbol{\beta}_i + \mathbf{u}_{ia}$, where \mathbf{u}_{ia} is a vector of domain specific random deviations from the fixed effects vector $\boldsymbol{\beta}_i$.

Non-linear GREG’s assisted by models such as (3.5) and (3.6) require model fitting for every y -variable separately; there is no uniformly applicable weight system. However, the question arises: Are there examples of non-linear GREG’s such that the practical advantages of linear GREG are preserved, that is, a linearly weighted form with calibrated weights independent of the y -variable. The answer is in the affirmative. Two directions in recent literature are of interest in this regard:

Breidt and Opsomer (2000), Montanari and Ranalli (2005) consider model-assisted local polynomial GREG estimators, for the case of a single continuous auxiliary variable with values x_k known for all $k \in U$. Several choices have to be made in the process: (1) the order q of the local polynomial expression, (2) the specification of the kernel function, and (3) the value of the band width. The resulting estimator can be expressed in terms of weights calibrated with respect to population totals of the powers of x_k , so that $\sum_s w_k x_k^j = \sum_U x_k^j$ for $j = 0, 1, \dots, q$.

Breidt, Claeskens and Opsomer (2005) develop a penalized spline GREG estimator for a single x -variable; the assisting model is $m(\mathbf{x}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \dots + \beta_q x^q + \sum_{j=1}^K \beta_{q+j} (x - \kappa_j)_+^q$, where $(t)_+^q = t^q$ if $t > 0$ and 0 otherwise, q is the degree of the spline, and the κ_j are suitably spaced knots, for example, uniformly spaced

sample quantiles of the x_k -values. After estimation of the β -parameters, they obtain the predictions $\hat{y}_k = m(x_k; \hat{\beta})$ needed for the general GREG formula (3.1). The authors point out that the resulting GREG estimator is calibrated for the parametric portion of the model, that is, $\sum_s w_k x_k^j = \sum_U x_k^j$ for $j = 0, 1, \dots, q$, and also for the truncated polynomial terms in the model as long as they are left unpenalized.

We can summarize GREG estimation as follows. The linear GREG has practical advantages for large scale statistics production: It can be expressed as a linearly weighted sum of y_k -values with weights calibrated to $\sum_U x_k$, the weights are independent of the y_k -values and may be applied to all y -variables in the survey. It is sufficient to know a population auxiliary total $\sum_U x_k$, imported from a reliable source. Non-linear GREG may give a considerably reduced variance, as a result of the more refined models that can be considered when there is complete auxiliary information (known x_k for all $k \in U$); near design unbiasedness is preserved. Certain non-linear GREG's can be written as linearly weighted sums.

In academic exercises with artificially created populations and relationships, one can provoke situations where a nonlinear GREG has a large variance advantage over a linear GREG. Such experiments are important for illustration. However, to meet the daily production needs in national statistical agencies; "farfetched" nonlinear GREG's seem to be of fairly remote interest at this point in time; the assisting models for GREG must meet requirements of robustness and practicality. The attraction of a minor reduction of the sampling variance is swept away by worries about other (non-sampling) errors and troubles in the daily production process.

The progression from linear to non-linear GREG creates opportunities and generates questions. What is the most appropriate formulation of the model expectation μ_k ? How sensitive are the results to the specification of the variance part of the assisting model? To what extent is computational efficiency an issue? Further research will respond more fully to these questions.

4. The calibration approach to estimation

4.1 Calibration under basic conditions

A crucial step in the GREG approach reviewed in the previous section is to produce predicted values \hat{y}_k through the fit of an assisting model. By contrast, the calibration approach, as defined in Section 1.1, does not refer explicitly to any model. It emphasizes instead the information on which one can calibrate. A key element of "calibration thinking" is the linear weighting of the observed y -values,

with weights made to confirm computable aggregates. This conceptual difference will sometimes lead to different estimators in the two approaches.

The calibration approach has considerable generality; it can deal with a variety of conditions: complex sampling designs, adjustments for nonresponse and frame errors. This section, however, focuses on the basic conditions in Section 2: single phase sampling and full response. The notation remains as in Section 2. The material available for estimating the population total $Y = \sum_U y_k$ is: (i) the study variable values y_k observed for $k \in s$, (ii) the known design weights $d_k = 1/\pi_k$ for $k \in U$, and (iii) the known vector values x_k for $k \in U$ (or an imported total $\sum_U x_k$). These simple conditions prevail in Deville and Särndal (1992) and Deville, Särndal and Sautory (1993), papers which gave the approach a name and inspired further work. Even though the background is simple, calibration raises several issues, some of them computational, as reviewed in Section 5.

The objective in Sections 4.2 and 4.3 is to determine weights w_k to satisfy the calibration equation $\sum_s w_k x_k = \sum_U x_k$, then use them to form the calibration estimator of Y as $\hat{Y}_{CAL} = \sum_s w_k y_k$, which we can confront with the unbiased Horvitz-Thompson estimator by writing $\hat{Y}_{CAL} = \hat{Y}_{HT} + \sum_s (w_k - d_k) y_k$. It follows that the bias of \hat{Y}_{CAL} is $E(\hat{Y}_{CAL}) - Y = E(\sum_s (w_k - d_k) y_k)$. Meeting the objective of near design unbiasedness requires $E(\sum_s (w_k - d_k) y_k) \approx 0$, whatever the y -variable. Evidently, the calibration should strive for small deviations $w_k - d_k$.

The objective "calibration for consistency with known population auxiliary totals" can be realized in many ways. We can construct many sets of weights calibrated to the known $\sum_U x_k$. This section examines this proliferation from two perspectives noted in the literature: the *minimum distance method* and the *instrumental vector method*. Yet another construction of a variety of calibrated weights is proposed in Demnati and Rao (2004).

4.2 The minimum distance method

In this method, the calibration sets out to modify the initial weights $d_k = 1/\pi_k$ into new weights w_k , determined to "be close to" the d_k . To this end, consider the distance function $G_k(w, d)$, defined for every $w > 0$, such that $G_k(w, d) \geq 0$, $G_k(d, d) = 0$, differentiable with respect to w , strictly convex, with continuous derivative $g_k(w, d) = \partial G_k(w, d) / \partial w$ such that $g_k(d, d) = 0$. Usually the distance function is chosen such that $g_k(w, d) = g(w/d) / q_k$, where the q_k are suitably chosen positive scale factors, $g(\cdot)$ is a function of a single argument, continuous, strictly increasing, with $g(1) = 0$, $g'(1) = 1$. Let $F(u) = g^{-1}(u)$ be the inverse function of $g(\cdot)$. Minimizing the total distance

$\sum_s G_k(w_k, d_k)$ subject to the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ leads to $w_k = d_k F(q_k \mathbf{x}'_k \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is obtained as the solution (assuming one exists) of

$$\sum_s d_k \mathbf{x}_k F(q_k \mathbf{x}'_k \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k. \quad (4.1)$$

The weights have an optimality property, because a duly specified objective function is minimized, but it is a “weak optimality” in the sense that there are many possible specifications of the distance function and the scale factors q_k .

Much attention has focused on the distance function $G_k(w_k, d_k) = (w_k - d_k)^2 / 2d_k q_k$. It gives $g_k(w_k, d_k) = (w_k/d_k - 1)/q_k$; $g(w/d) = w/d - 1$; $F(u) = g^{-1}(u) = 1 + u$. The term “the linear case” is thus appropriate. The task is then to minimize the “chi-square distance” $\sum_s (w_k - d_k)^2 / 2d_k q_k$, subject to $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Equation (4.1) reads $\sum_s d_k \mathbf{x}_k (1 + q_k \mathbf{x}'_k \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$, which is easily solved for $\boldsymbol{\lambda}$. The resulting estimator of $Y = \sum_U y_k$ is $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights $w_k = d_k g_k$ given by (3.3). That is, $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}}$ as given by (3.2), and the residuals that determine the asymptotic variance are $E_k = y_k - \mathbf{x}'_k \mathbf{B}_{U,q}$ as given in Section 3.2. Some negative weights w_k may occur.

The linear GREG estimator implies weights that happen to be calibrated (to $\sum_U \mathbf{x}_k$), and the opposite side of the same coin says that the linear case for calibration (with chi-square distance) brings the linear GREG estimator. The tendency in some articles and applications to intertwine GREG thinking and calibration thinking stems from this fact. Many successful applications of the use of auxiliary information stem, in any case, from this linearity on both sides of the coin. The Canadian Labour Force Survey is an example, and an interesting recent development for that survey is the use of composite estimators, with part of the information coming from the survey results in previous months, as described in Fuller and Rao (2001).

The calibration equation is satisfied for any choice of the positive scale factors q_k in (4.1). A simple choice is $q_k = 1$ for all k . But it is not always the preferred choice. For example, if there is a single, always positive auxiliary variable, and $\mathbf{x}_k = x_k$, then many will intuitively expect $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ to deliver the usual ratio estimator $\sum_U x_k (\sum_s d_k y_k) / (\sum_s d_k x_k)$, and it does, but by taking $q_k = x_k^{-1}$, not $q_k = 1$.

Another distance function of considerable interest is $G_k(w_k, d_k) = \{w_k \log(w_k/d_k) - w_k + d_k\} / q_k$. It leads to $F(u) = g^{-1}(u) = \exp(u)$, “the exponential case”. Then (4.1) reads $\sum_s d_k \mathbf{x}_k \exp(q_k \mathbf{x}'_k \boldsymbol{\lambda}) = \sum_U \mathbf{x}_k$. Numeric methods are required to solve for $\boldsymbol{\lambda}$, to obtain the weights $w_k = d_k \exp(q_k \mathbf{x}'_k \boldsymbol{\lambda})$. No negative weights w_k will occur.

Deville and Särndal (1992) show that a variety of distance functions satisfying mild conditions will generate

asymptotically equivalent calibration estimators. Alternative distance functions are compared in Deville, Särndal and Sautory (1993), Singh and Mohl (1996), Stukel, Hidiroglou and Särndal (1996). Some distance functions will guarantee weights falling within specified bounds, so as to rule out too large or too small (negative) weights. Changes in the distance function will often have minor effect only on the variance of the calibration estimator $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$, even if the sample size is rather small. Questions about the existence of a solution to the calibration equation are discussed in Th  berge (2000).

4.3 The instrument vector method

An alternative to distance minimization is the instrumental vector method, considered in Deville (1998), Estevao and S  r  ndal (2000, 2006) and Kott (2006). It can also generate many alternative sets of weights calibrated to the same information.

We can consider weights of the form $w_k = d_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$, where \mathbf{z}_k is a vector with values defined for $k \in s$ and sharing the dimension of the specified auxiliary vector \mathbf{x}_k , and the vector $\boldsymbol{\lambda}$ is determined from the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. The function $F(\cdot)$ plays the same role as in the distance minimization method; several choices $F(\cdot)$ are of interest, for example, $F(u) = 1 + u$ and $F(u) = \exp(u)$.

Opting for the linear function $F(u) = 1 + u$, we have $w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k)$. It is an easy exercise to determine $\boldsymbol{\lambda}$ to satisfy the calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. The resulting calibration estimator is

$$\hat{Y}_{\text{CAL}} = \sum_s w_k y_k; w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k),$$

$$\boldsymbol{\lambda}' = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)' \left(\sum_s d_k \mathbf{z}_k \mathbf{x}'_k \right)^{-1}. \quad (4.2)$$

Whatever the choice of \mathbf{z}_k , the weights $w_k = d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k)$ satisfy the calibration equation. The standard choice is $\mathbf{z}_k = \mathbf{x}_k$. In particular, setting $\mathbf{z}_k = q_k \mathbf{x}_k$, for specified q_k , gives the weights (3.3).

Even “deliberately awkward choices” for \mathbf{z}_k give surprisingly good results. For example, let x_k be a single continuous auxiliary variable, and $\mathbf{z}_k = c_k x_k^{p-1}$. Suppose $p = 3$, and $c_k = 1$ for 4 elements only, chosen at random from $n = 100$ elements in a realized sample s , and $c_k = 0$ for the remaining 96. The near-unbiasedness of $\hat{Y}_{\text{CAL}} = \sum_s d_k (1 + \boldsymbol{\lambda}' \mathbf{z}_k) y_k$ is still present. Even with such a sparse \mathbf{z} -vector, the increase in variance, relative to better choices of \mathbf{z}_k , may not be excessive.

When both sampling design and \mathbf{x} -vector are fixed, Estevao and S  r  ndal (2004) and Kott (2004) note that there is an asymptotically optimal \mathbf{z} -vector given by

$$\mathbf{z}_k = \mathbf{z}_{0k} = d_k^{-1} \sum_{\ell \in s} (d_k d_\ell - d_{k\ell}) \mathbf{x}_\ell$$

where $d_{k\ell}$ is the inverse of the second order inclusion probability $\pi_{k\ell} = P(k \& \ell \in s)$, assumed strictly positive. The resulting calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_s d_k (1 + \lambda' z_{0k}) y_k$, is essentially the “randomization-optimal estimator” due originally to Montanari (1987) and discussed by many since then.

Andersson and Thorburn (2005) view the question from the opposite direction and ask: In the minimum distance method, can a distance function be specified such that its minimization will deliver the randomization-optimal estimator? They do find this distance; not entirely surprisingly, it is related to (but not identical to) the chi-square distance.

4.4 Does calibration need an explicitly stated model?

The calibration approach as presented in Sections 4.2 and 4.3 proceeds by simply computing the weights that reproduce the specified auxiliary totals. There is no explicit assisting model, unless one were to insist that picking certain variables for inclusion in the vector \mathbf{x}_k amounts to a serious modeling effort. Instead, the weights are justified primarily by their consistency with the stated controls. Early contributions reflect this attitude, from Deming (1943), and continuing with Alexander (1987), Zieschang (1990) and others. This begs the question: Is it nevertheless important to motivate such “model-free calibration” with an explicit model statement? It is true that statisticians are trained to think in terms of models, and they feel more or less compelled to always have a statistical procedure accompanied by a model statement. It may indeed have some pedagogical merit, also in explaining calibration, to state the associated relationship of y to \mathbf{x} , even if it is as simple as a standard linear model.

But will a stated model help the users and practitioners better understand the calibration approach? To most of them the approach is perfectly clear and transparent anyway. They need no other justification than the consistency with stated controls. Will a search for “the true model with the true variance structure” bring significantly better accuracy for the bulk of the many estimates produced in a large government survey? It is unlikely.

The next section deals with model-calibration. For that variety, proposed by Wu and Sitter (2001), modeling has indeed an explicit and prominent role. These authors call the linear calibration estimator, $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights w_k given by (3.3), “a routine application without modeling”. The description is appropriate in that all that is necessary is to identify the x -variables with their known population totals.

4.5 Model-calibration

The idea of model-calibration is proposed in Wu and Sitter (2001) and pursued further in Wu (2003) and

Montanari and Ranalli (2003, 2005). The motivating factor is that complete auxiliary information allows a more effective use of the \mathbf{x}_k known for every $k \in U$ than what is possible in model-free calibration, where a known total $\sum_U \mathbf{x}_k$ is sufficient. The weights are required to be consistent with the computable population total of the predictions \hat{y}_k , derived via an appropriate model formulation. Thus the weight system may not be consistent with the known population total of each auxiliary variable, unless there is special provision to retain this property. Model-calibration still satisfies all three parts, (a) to (c), of the definition of calibration in Section 1.1; in particular, the estimators are nearly design unbiased.

Consider a non-linear assisting model of the type (3.4). We estimate the unknown parameter θ by $\hat{\theta}$, leading to fitted values $\hat{y}_k = \hat{\mu}_k = \mu(\mathbf{x}_k, \hat{\theta})$ computed with the aid of the \mathbf{x}_k known for all $k \in U$. It follows that the population size N is known and should be brought to play a significant role in the calibration. If minimum chi-square distance is used, we find the weights of the model-calibration estimator $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$ by minimizing $\sum_s (w_k - d_k)^2 / (2d_k q_k)$, for specified q_k , and $d_k = 1/\pi_k$, subject to the calibration equations

$$\sum_s w_k = N; \sum_s w_k \hat{y}_k = \sum_U \hat{y}_k. \quad (4.3)$$

For simplicity, let us take $q_k = 1$ for all k ; we derive the calibrated weights, rearrange them and find that the model-calibration estimator can be written as

$$\hat{Y}_{\text{MCAL}} = N \{ \bar{y}_{s;d} + (\bar{y}_U - \bar{y}_{s;d}) \tilde{B}_{s;d} \} \quad (4.4)$$

where $\bar{y}_{s;d} = \sum_s d_k y_k / \sum_s d_k$; $\bar{y}_{s;d} = \sum_s d_k \hat{y}_k / \sum_s d_k$, and

$$\tilde{B}_{s;d} = \left(\sum_s d_k (\hat{y}_k - \bar{y}_{s;d}) y_k \right) / \sum_s d_k (\hat{y}_k - \bar{y}_{s;d})^2.$$

The regression implied by $\tilde{B}_{s;d}$ is one of observed y -values on predicted y -values. The idea of this regression would hardly occur to the modeler is his/her attempts to structure the relation between y_k and \mathbf{x}_k , but it proves effective in building the calibration estimator. Wu and Sitter (2001) present evidence that

$$(\hat{Y}_{\text{MCAL}} - Y) / N = \left(\sum_s d_k \tilde{E}_k - \sum_U \tilde{E}_k \right) / N + O_p(n^{-1})$$

with $\tilde{E}_k = y_k - \bar{y}_U - (\mu_k - \bar{\mu}_U) \tilde{B}_U$, where $\tilde{B}_U = (\sum_U (\mu_k - \bar{\mu}_U) y_k) / (\sum_U (\mu_k - \bar{\mu}_U)^2)$, and $\bar{\mu}_U = \sum_U \mu_k / N$. The coefficient \tilde{B}_U may not be near one even in large samples. It expresses a regression of y_k on its assisting model mean $\mu_k = \mu(\mathbf{x}_k, \beta)$. That is, \hat{Y}_{MCAL} can be viewed as a regression estimator that uses the model expectation μ_k as the auxiliary variable, leaving \tilde{E}_k as the residuals that determine the asymptotic variance of \hat{Y}_{MCAL} .

How does this asymptotic variance compare with that of the non-linear GREG construction (3.1) for the same non-linear assisting model and the same $\hat{y}_k = \hat{\mu}_k$? Formula (3.1) implies a slope equal to unity in the regression between y_k and $\hat{y}_k = \hat{\mu}_k$; viewed in that light, \hat{Y}_{GREG} is a difference estimator rather than a regression estimator and hence less sensitive to the pattern in the data. The non-linear GREG \hat{Y}_{GREG} is in general less efficient than \hat{Y}_{MCAL} . (It is of course possible to modify \hat{Y}_{GREG} to also account for the information contained in the known population size N .)

On the other hand, compared with the linear (model-free) calibration estimator $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ with weights as in (3.3), the model-calibration estimator \hat{Y}_{MCAL} given by (4.4) may have a considerable variance advantage but implies a loss of the practical advantages of a consistency with the known population total $\sum_U x_k$ and a multi-purpose weight system applicable to all y -variables. The y -values in (4.4) are linearly weighted, but the weights now also depend on the y -values. It is thus debatable if \hat{Y}_{MCAL} is a bona fide calibration estimator.

In an empirical study, Wu and Sitter (2001) compare $\hat{Y}_{\text{MCAL}} = \sum_s w_k y_k$, calibrated according to (4.3), with the non-linear GREG, $\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_k + \sum_s d_k (y_k - \hat{y}_k)$ given by (3.1), for the same non-linear assisting model and same $\hat{y}_k = \hat{\mu}_k$. The study confirms that \hat{Y}_{MCAL} has a variance advantage over the non-linear \hat{Y}_{GREG} . They created a finite population U of size $N = 2,000$ with values (y_k, x_k) , $k = 1, \dots, 2,000$, such that $\log(y_k) = 1 + x_k + \varepsilon_k$; the 2,000 values x_k are realizations of the Gamma(1,1) random variable, and ε_k is a normally distributed error. The auxiliary information consists of the population size N and the known values x_k for $k = 1, \dots, 2,000$. Repeated simple random samples of size $n = 100$ were taken; the assisting model for both estimators was the log-linear $E_\varepsilon(y_k | x_k) = \mu_k$ with $\log(\mu_k) = \alpha + \beta x_k$. This model was fit for each sample, using pseudo-maximum quasi-likelihood estimation. The fitted values $\hat{y}_k = \exp(\hat{\alpha} + \hat{\beta} x_k)$ were used to form both \hat{Y}_{MCAL} and \hat{Y}_{GREG} . The simulation variance was markedly lower for \hat{Y}_{MCAL} . (The linear GREG (3.2), identical to the model-free calibration estimator, was also included in the Wu and Sitter study; not surprisingly, it is even less efficient than the non-linear GREG, under the strongly non-linear relationship imposed in their experiment.)

Montanari and Ranalli (2005) provide further evidence, for several artificially created populations, on the comparison between \hat{Y}_{MCAL} and the non-linear \hat{Y}_{GREG} . Their assisting model, $y_k = \mu_k + \varepsilon_k$, is fitted via nonparametric regression (local polynomial smoothing), yielding predictions $\hat{y}_k = \hat{\mu}_k$ for $k \in U$. With this type of model fit, the predictions $\hat{y}_k = \hat{\mu}_k$ are highly accurate. Not surprisingly, the model-calibration estimator \hat{Y}_{MCAL}

achieves only marginal improvement over the non-linear \hat{Y}_{GREG} .

We can summarize the calibration approach as follows: The estimator of $Y = \sum_U y_k$ has the linearly weighted form $\hat{Y} = \sum_s w_k y_k$. In linear (model-free) calibration, the calibration equation reads $\sum_s w_k x_k = \sum_U x_k$; a known population auxiliary total $\sum_U x_k$ is required, but complete auxiliary information (known x_k for all $k \in U$) is not required; the same weights can be applied to all y -variables (multi-purpose weighting); the estimator is identical to the linear GREG estimator (but derived by different reasoning). In model-calibration, the assisting model mean μ_k is non-linear in x_k ; complete auxiliary information is usually required; the calibration constraints include the equation $\sum_s w_k \hat{y}_k = \sum_U \hat{y}_k$; the weights w_k depend on the y_k -values, implying a loss of the multi-purpose property.

5. Computational aspects, extreme weights and outliers

The computation of calibrated weights raises important practical issues, discussed in a number of papers. All computation must proceed smoothly and routinely in the large scale statistics production of a national statistical agency. Undesirable (or unduly variable) weights should be avoided. Many practitioners support the reasonable requirement that all weights be positive (even greater than unity) and that very large weights should be avoided.

A few of the weights computed according to (3.2) can turn out to be quite large or negative. Huang and Fuller (1978) and Park and Fuller (2005) proposed methods to avoid undesirable weights.

In the distance minimization method, the distance function can be formulated so that negative weights are excluded, while still satisfying the given calibration equations. The software CALMAR (Deville, Särndal and Sautory 1993) allows several distance functions of this kind. An expended version, CALMAR2, is described in LeGuennec and Sautory (2002). Other statistical agencies have developed their own software for weight computation. Among those are GES (Statistics Canada), CLAN97 (Statistics Sweden), Bascula 4.0 (Central Bureau of Statistics, The Netherlands), g-CALIB-S (Statistics Belgium). These strive, in different ways, to resolve the computational issues arising. The user needs to consult the users' guide in each particular case to see exactly how the computational issues, including an avoidance of undesirable weights, are handled.

GES uses mathematical programming to minimize the chi-square distance, subject to the calibration constraints as well as to individual bounds on the weights, so that they will satisfy $A_k \leq w_k \leq B_k$ for specified A_k, B_k . Bascula 4.0 is

described in Nieuwenbroek and Boonstra (2002). The software g-CALIB-S, described in Vanderhoeft, Waeytens and Museux (2001), Vanderhoeft (2001), uses generalized inverse (the Moore–Penrose) for the weight computation; consequently one need not be concerned about a possible redundancy in the auxiliary information.

In Bankier, Houle and Luc (1997) the objective is two-fold: to keep the computed weights within desirable bounds, and to drop some x -variables to remove near-linear dependencies. Isaki, Tsay and Fuller (2004) consider quadratic programming to obtain both household weights and person weights that lie within specified bounds.

An intervention with the weights (so as to get rid of undesirable weight values) raises the question how far one can deviate from the design weights d_k without compromising the desirable feature of nearly design unbiased estimation. An idea that has been tried is to modify the set of constraints so that tolerances are respected for the difference between the estimator for the auxiliary variables and the corresponding known population totals. Hence, Chambers (1996) minimizes a “cost-ridged loss function”.

Outlying values in the auxiliary variables may be a cause of extreme weights. Calibration in the presence of outliers is discussed in Duchesne (1999). His technique of “robust calibration” may introduce a certain bias in the estimates; it may, however, be more than offset by a reduction in variance.

When the set of constraints is extended to make the weights restricted to specified intervals, a solution to the optimization problem is not guaranteed. The existence of a solution is considered in Théberge (2000), who also proposes methods for dealing with outliers.

6. Calibration estimation for more complex parameters

The calibration approach adapts itself to the estimation of more complex parameters than a population total. Examples are reviewed in this section. Single phase sampling and full response continue to be assumed; the notation remains as in Section 2. One example is the estimation of population quantiles (Section 6.1), another is the estimation of functions of totals (Section 6.2). Other examples in this category, not reviewed here, are Théberge (1999), for the estimation of bilinear parameters, and Tracy, Singh and Arnab (2003), for calibration with respect to second order moments.

6.1 Calibration for estimation of quantiles

The median and other quantiles of the finite population are important descriptive measures, especially in economic surveys. To estimate quantiles, the finite population

distribution function must first be estimated. Before calibration became popular, several papers considered the estimation of quantiles, with or without the use of auxiliary information. More recent articles have turned to the calibration approach for the same purpose, including Kovačević (1997), Wu and Sitter (2001), Ren (2002), Tillé (2002), Harms (2003), Harms and Duchesne (2006) and Rueda *et al.* (2007). As these papers illustrate, there is more than one way to implement the calibration approach. The non-smooth character of the finite population distribution function causes certain complexities; these are resolved by different authors in different ways.

Let $\Delta(\cdot)$ denote the Heaviside function, defined for all real z so that $\Delta(z) = 1$ if $z \geq 0$ and $\Delta(z) = 0$ if $z < 0$. The unknown distribution function of the study variable y is

$$F_y(t) = \frac{1}{N} \sum_U \Delta(t - y_k). \quad (6.1)$$

The α -quantile of the finite population is defined as $Q_{y\alpha} = \inf\{t | F_y(t) \geq \alpha\}$. The auxiliary variable x_j , taking values x_{jk} , has the distribution function $F_{x_j}(t) = (1/N) \sum_U \Delta(t - x_{jk})$ with α -quantile denoted $Q_{x_j\alpha}$, $j = 1, 2, \dots, J$. A natural estimator of $F_y(t)$ based on the design weights $d_k = 1/\pi_k$ is

$$\hat{F}_y(t) = \frac{1}{\sum_s d_k} \sum_s d_k \Delta(t - y_k).$$

A calibration estimator $F_{y\alpha}(t)$ of takes the form

$$\hat{F}_{y\text{CAL}}(t) = \frac{1}{\sum_s w_k} \sum_s w_k \Delta(t - y_k) \quad (6.2)$$

where the weights w_k are suitably calibrated to a specified auxiliary information; then from $\hat{F}_{y\text{CAL}}(t)$ we obtain the α -quantile estimator as $\hat{Q}_{y\alpha} = \inf\{t | \hat{F}_{y\text{CAL}}(t) \geq \alpha\}$. A formula analogous to (6.2) holds for $\hat{F}_{x_j\text{CAL}}(t)$.

Without explicit reference to any model, Harms and Duchesne (2006) specify the information available for calibration as a known population size, N , and known population quantiles $Q_{x_j\alpha}$ for $j = 1, 2, \dots, J$. The complete auxiliary information, with values $\mathbf{x}_k = (x_{k1}, \dots, x_{kJ})'$ known for $k \in U$, is not required. (But in practice, the complete information would usually be necessary, because accurate quantiles of several x -variables are not likely to be importable from outside sources.) They determine the w_k to minimize the chi-square distance $\sum_s (w_k - d_k)^2 / 2d_k q_k$, for specified q_k , subject to the calibration equations

$$\sum_s w_k = N; \hat{Q}_{x_j\text{CAL}, \alpha} = Q_{x_j\alpha}, j = 1, 2, \dots, J$$

for suitably defined estimates $\hat{Q}_{x_j\text{CAL}, \alpha}$. Now, if we were to specify $\hat{Q}_{x_j\text{CAL}, \alpha} = \inf\{t | \hat{F}_{x_j\text{CAL}}(t) \geq \alpha\}$, then it is in general not possible to find an exact solution of the calibration

problem as stated. Instead, Harms and Duchesne substitute smoothed estimators, called “interpolated distribution estimators”, of the distribution functions $F_{x_j}(t)$, $j=1, 2, \dots, J$. They replace $\Delta(\cdot)$ by a slightly modified function. Weights w_k can now be obtained, as well as a corresponding estimated distribution function $\hat{F}_{yCAL}(t)$; finally, $Q_{y\alpha}$ is estimated as $\hat{Q}_{y\alpha} = \hat{F}_{yCAL}^{-1}(\alpha)$.

The resulting calibrated weights w_k allow us to retrieve the known population quantiles of the auxiliary variables. This is reassuring; one would expect such weights to produce reasonable estimators for the quantiles of the study variable y . Moreover, in the case of a single scalar auxiliary variable x , the resulting calibration estimator delivers exact population quantiles for y when the relationship between y and x is exactly linear, that is, when $y_k = \beta x_k$ for all $k \in U$. An idea involving smoothed distribution functions is also used in Tillé (2002).

The computationally simpler method of Rueda *et al.* (2007) is an application of model-calibration, in that they calibrate with respect to a population total of *predicted* y -values. Complete auxiliary information is required. Using the known \mathbf{x}_k , compute first the linear predictions $\hat{y}_k = \hat{\beta}'\mathbf{x}_k$ for $k \in U$, with $\hat{\beta} = (\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_s d_k q_k \mathbf{x}_k y_k)$, where $d_k = 1/\pi_k$ and the q_k are specified scale factors. The weights w_k are obtained by minimizing the chi-square distance subject to calibration equations stated in terms of the predictions, so as to have consistency at J arbitrarily chosen points t_j , $j=1, \dots, J$:

$$\frac{1}{N} \sum_s w_k \Delta(t_j - \hat{y}_k) = F_y(t_j), j=1, \dots, J$$

where $F_y(t_j)$ is the finite population distribution function of the predictions \hat{y}_k , evaluated at t_j . It is suggested that a fairly small number of arbitrarily selected points t_j may suffice, say less than 10. Once the w_k are determined, the α -quantile estimate is obtained from $\hat{F}_{yCAL}(t) = (1/N) \sum_s w_k \Delta(t - y_k)$.

Quantile estimation provides a good illustration that the calibration approach can be carried out in more than one way when somewhat more complex parameters are being estimated. Both methods mentioned give nearly design unbiased estimation. The Harms and Duchesne (2006) weights are multi-purpose, independent of the y -variable; by contrast, the method of Rueda *et al.* (2007) requires a new set of weights for every new y -variable. Empirical evidence, by simulation, suggests that both methods compare favourably with the earlier quantile estimation methods, not based explicitly on calibration thinking (but on the same auxiliary information).

An extension of the calibration approach to the estimation of other complex parameters, such as the Gini coefficient, is sketched in Harms and Duchesne (2006).

6.2 Calibration for other complex parameters

Plikusas (2006), and Krapavickaitė and Plikusas (2005) examine calibration estimation of certain functions of population totals. (Their term “non-linear calibration” signifies “non-linear function of totals”; I do not use it here.) A simple example is the estimation of a ratio of two totals, $R = \sum_U y_{1k} / \sum_U y_{2k}$, where y_{1k} and y_{2k} are the values for element k of the variables y_1 and y_2 , respectively. (The distribution function (6.1) is in effect also of ratio type, with $y_{2k} = 1$, and $N = \sum_U 1$ as the denominator total.) These authors examine the calibration estimator $\hat{R}_{CAL} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Its weights w_k , common to the numerator and the denominator, are determined by calibration to auxiliary information stated as follows: There is one auxiliary variable, x_{1k} , for y_{1k} , and another, x_{2k} , for y_{2k} ; the ratio of totals $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ is a known value, by a complete enumeration at a previous occasion or from some other accurate source. The proposed calibration equation is $\sum_s w_k e_k = 0$, where $e_k = x_{1k} - R_0 x_{2k}$. Because $\sum_U e_k = 0$, the weights, by minimum chi-square distance, are

$$w_k = d_k \left\{ 1 - \left(\sum_s d_k e_k \right) \left(\sum_s d_k e_k^2 \right)^{-1} e_k \right\}.$$

These weights correctly retrieve the known ratio value R_0 ; setting $y_{1k} = x_{1k}$ and $y_{2k} = x_{2k}$ in \hat{R}_{CAL} , we have

$$\frac{\sum_s w_k x_{1k}}{\sum_s w_k x_{2k}} - R_0 = \frac{\sum_s w_k e_k}{\sum_s w_k x_{2k}} = 0.$$

The empirical evidence in Plikusas (2006), and Krapavickaitė and Plikusas (2005) suggests that their calibration estimator compares favourably (lower variance, while maintaining near design unbiasedness) with other estimators, derived through other arguments than calibration, while relying on the same auxiliary information.

7. Calibration contrasted with other approaches

As many have noted, users view calibration as a simple and convincing way to incorporating auxiliary information, for simple parameters (Section 4), as for more complex parameters such as quantiles, ratios and others (Section 6). Simplicity and practicality are undeniable advantages, but aside from that, is calibration also “theoretically superior”? Are there instances where calibration can be shown to give more accurate and/or more satisfactory answers on questions of importance, when contrasted with other design-based approaches?

Section 4.5 gave one indication that calibration thinking may have an advantage over GREG thinking, in that model-calibration may give more precise estimates than the

non-linear GREG, for the same assisting model. The following Section 7.1 gives another example where calibration reasoning and GREG reasoning give diverging answers, with an advantage for the calibration method.

7.1 An example in domain estimation

The example in this section, from Estevao and Särndal (2004), shows, for a simple practical situation, a conflict between the results of GREG thinking and calibration thinking. The context is the estimation of the y -total for a sub-population (a domain).

A probability sample s is drawn from $U = \{1, 2, \dots, k, \dots, N\}$; the known design weights are $d_k = 1/\pi_k$. Let U_a be a domain; $U_a \subset U$. The domain indicator is δ_{ak} with value $\delta_{ak} = 1$ if $k \in U_a$ and $\delta_{ak} = 0$ if not. The target of estimation is the domain total $Y_a = \sum_U y_{ak}$, where $y_{ak} = \delta_{ak} y_k$, and y_k is observed for $k \in s$. The Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_s d_k y_{ak}$, although design unbiased, has low precision, especially if the domain is small; the use of auxiliary information will bring improvement. An auxiliary vector value \mathbf{x}_k is specified for every $k \in U$.

As is frequently the case in practice, the elements belonging to a domain of interest are not identified in the sampling frame. (If they are, some very powerful information is available from the start, but frequently real world conditions are not that favourable.) But suppose elements in a larger group U_C are identifiable; $U_a \subset U_C \subset U$. For example, suppose y is "income" and U_C a professional group specified for the persons listed in the frame, while U_a is a professional sub-group not identified in the frame. We can identify the sample subsets $s_C = s \cap U_C$ and $s_a = s \cap U_a$, and we can benefit from knowing the total $\sum_U \mathbf{x}_{Ck}$, estimable without bias by $\sum_s d_k \mathbf{x}_{Ck}$, where $\mathbf{x}_{Ck} = \delta_{Ck} \mathbf{x}_k$, and δ_{Ck} is the information group indicator: $\delta_{Ck} = 1$ if $k \in U_C$ and $\delta_{Ck} = 0$ if not. The domain auxiliary total $\sum_U \mathbf{x}_{ak}$ is unavailable, because U_a is not identified. Calibration to satisfy $\sum_s w_k \mathbf{x}_{Ck} = \sum_U \mathbf{x}_{Ck}$ gives the nearly design unbiased estimator $\hat{Y}_{CAL} = \sum_U w_k y_{ak}$, where $w_k = d_k (1 + \lambda' \mathbf{z}_k)$, with $\lambda' = (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck})' (\sum_s d_k \mathbf{x}_k \mathbf{x}_{Ck}')^{-1}$. The asymptotically optimal instrument for the given vector \mathbf{x}_k is (see Section 4.3) $\mathbf{z}_k = \mathbf{z}_{0Ck} = d_k^{-1} \sum_{\ell \in s} (d_k d_\ell - d_{k\ell}) \mathbf{x}_{C\ell}$.

By contrast, regression thinking for the same auxiliary information leads to $\hat{Y}_{GREG} = \sum_s d_k y_{ak} + (\sum_U \mathbf{x}_{Ck} - \sum_s d_k \mathbf{x}_{Ck}) \mathbf{B}_{\tilde{s};d}$, also nearly design unbiased, where the regression coefficient $\mathbf{B}_{\tilde{s};d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s d_k \mathbf{x}_k y_k$ is the result of a weighted least squares fit at a suitable level, using all (when $\tilde{s} = s$) or part (when $\tilde{s} \subset s$) of the data points (y_k, \mathbf{x}_k) available for $k \in s$.

For example, the modeller may opt for a regression fit "extending beyond the domain" (so that $\tilde{s} \supset s_a = s \cap U_a$),

in an attempt to borrow strength for \hat{Y}_{aGREG} by letting it depend also on y -data from outside the domain. By contrast, \hat{Y}_{qCAL} relies exclusively on y -data in the domain, and this is in effect better. Estevao and Särndal (2004) show that \hat{Y}_{aCAL} with $\mathbf{z}_k = \mathbf{z}_{0Ck}$ has smaller (asymptotic) variance than \hat{Y}_{aGREG} , no matter how \tilde{s} is chosen. Bringing in y -data from the outside does not help; calibration thinking and regression thinking do not agree.

8. Calibration estimation in the presence of composite information

As the preceding sections have shown, many papers choose to study estimation for direct, single phase sampling of elements, without any nonresponse. The information available for calibration is simple; the k :th element of the finite population $U = \{1, 2, \dots, k, \dots, N\}$ has an associated auxiliary vector value \mathbf{x}_k .

However, in an important category of situations, the auxiliary information has *composite structure*. The complexity of the information increases with that of the sampling design. In designs with two or more phases, or in two or more stages, the information is typically composed of more than one component, reflecting the features of the design. The information is stated in terms of more than one auxiliary vector. For example, in two-stage sampling, some information may be available about the first stage sampling units (the clusters), other information about the second stage units (the elements).

Consequently, estimation by calibration (or by any alternative method) must take the composite structure of the information systematically into account. The total information has several pieces; the calibration can be done in more than one way. All relevant pieces should be taken into account, for best possible accuracy in the estimates. To accomplish this in a general or "optimal" way is not a trivial task. Calibration reasoning offers one way.

Regression reasoning, with a duly formulated assisting model, is an alternative way, but it will strike some users as more roundabout. Hence, surveys that allow composite auxiliary information bring further perspectives on the contrast between calibration thinking and GREG thinking.

Two-phase sampling and two-stage sampling are discussed in this section. Another example of composite information occurs for nonresponse bias adjustment, as discussed in Section 9.

Another aspect of composite information occurs when the objective is to combine information from several surveys. This, too, can be a way to add strength and improve accuracy of the estimates. It is a motivating factor (in addition to the user oriented motive to achieve consistency

among surveys) in the previously mentioned repeated weighting methodology of the Dutch statistical agency. Combined auxiliary information for GREG estimation is considered in Merkouris (2004).

8.1 Composite information for two-phase sampling designs

Double sampling refers to designs involving two probability samples, s_1 and s_2 , from the same population $U = \{1, \dots, k, \dots, N\}$. Auxiliary data may be recorded for both U and s_1 , the study variable values y_k are recorded only for $k \in s_2$ with an objective to estimate $Y = \sum_U y_k$. Hidiroglou (2001) distinguishes several kinds of double sampling: In the *nested case* (traditional two phase sampling), the first phase sample s_1 is drawn from U , the second phase sample s_2 is a sub-sample from s_1 , so that $U \supset s_1 \supset s_2$. Two *non-nested cases* can be distinguished: In the first of these, s_1 is drawn from the frame U_1 ; s_2 from the frame U_2 , where U_1 and U_2 cover the same population U ; the sampling units may be defined differently for the two frames. In the second non-nested case, s_1 and s_2 are drawn independently from U .

To illustrate how composite information intervenes in the estimation, consider the nested case. The design weights are $d_{1k} = 1/\pi_{1k}$ (s_1 sampled from U); $d_{2k} = 1/\pi_{2k}$ ($\pi_{2k} = \pi_{k|s_1}$ in sub-sampling s_2 from s_1). The combined design weight is $d_k = d_{1k}d_{2k}$. The basic unbiased estimator $\hat{Y} = \sum_{s_2} d_k y_k$ can be improved by a use of auxiliary information, specified here at two levels:

Population level: The vector value \mathbf{x}_{1k} is known (given in the frame) for every $k \in U$, thus known for every $k \in s_1$ and for every $k \in s_2$; $\sum_U \mathbf{x}_{1k}$ is a known population vector total;

First sample level: The vector value \mathbf{x}_{2k} is known (observed) for every $k \in s_1$, and thereby known for every $k \in s_2$; the unknown total $\sum_U \mathbf{x}_{2k}$ is estimated without bias by $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$.

How do we best take this composite information into account? In an adaptation of GREG thinking, Särndal and Swensson (1987) formulated two linear assisting models, the first one stated in terms of the \mathbf{x}_{1k} -vector, the other one also brings in the \mathbf{x}_{2k} -vector. The two models are fitted; the resulting predictions, of two kinds, are used to create an appropriate GREG estimator \hat{Y}_{GREG} of $Y = \sum_U y_k$.

Dupont (1995) makes the important point that the given composite information invites “two different natural approaches”: Besides the GREG approach, there is a calibration approach that will deliver final weights w_k for a calibration estimator $\hat{Y}_{\text{CAL}} = \sum_{s_2} w_k y_k$. It is of interest to compare the results of the two approaches. Both of them allow more than one option: In the GREG approach, there are alternative ways of formulating the linear assisting

models with their respective variance structures. In the calibration approach, alternative formulations of the calibration equations are possible.

For example, a *two-step calibration* option is as follows: First find intermediate weights w_{1k} to satisfy $\sum_{s_1} w_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$; then use the weights w_{1k} in the second step to compute the final weights w_k to satisfy

$$\sum_{s_2} w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k = \left(\sum_U \mathbf{x}_{1k} \right)$$

where \mathbf{x}_k is the combined auxiliary vector

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}.$$

Alternatively, in a *single step* option, we determine the w_k directly to satisfy

$$\sum_{s_2} w_k \mathbf{x}_k = \left(\sum_U \mathbf{x}_{1k} \right).$$

The final weights w_k are in general not identical in the two options. Suppose that $\sum_U \mathbf{x}_{1k}$ is an imported \mathbf{x}_1 -total. At closer look, the two-step option requires more extensive information, because individually known values \mathbf{x}_{1k} are required for $k \in s_1$, whereas it is sufficient in the single step option that they be available for $k \in s_2$. Some variance advantage may thus be expected from the two-step option, since $\sum_{s_1} w_{1k} \mathbf{x}_{2k}$ is often more accurate (as an estimator of $\sum_U \mathbf{x}_{2k}$) than $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$ in the single step procedure. Nevertheless, this anticipation is not always confirmed; the single step method can be better, as when \mathbf{x}_1 and \mathbf{x}_2 are weakly correlated.

Dupont (1995) and Hidiroglou and Särndal (1998) examine links that exist, not surprisingly, between the two approaches. A GREG estimator, derived from assisting models with specific variance structures, may be identical to calibration estimator, if the weights of the latter are calibrated in a certain way. In other cases, differences may be small.

The efficiency of different options depends in rather subtle ways on the pattern of correlation among y_k , \mathbf{x}_{1k} and \mathbf{x}_{2k} . For example, to what extent do \mathbf{x}_1 and \mathbf{x}_2 complement each other, to what extent are they substitutes for one another? In the GREG approach, it is difficult or even futile to pinpoint a variance structure that truly captures a “reality” behind the data. The calibration approach is more direct. Some of its possibilities are explored in Estevao and Särndal (2002, 2006).

8.2 Composite information in two-stage sampling designs

The traditional two-stage sampling set-up (clusters sampled at stage one, elements sub-sampled within selected

clusters in stage two) has in common with two-phase sampling that the total information may have more than one component. There may exist (a) information at the cluster level (about the clusters); (b) information at the element level for all clusters; (c) information at the element level for the selected clusters only. Here again, authors are of two different orientations: some exploit the information via calibration thinking, others follow the GREG thinking route.

Estevao and Särndal (2006) develop calibration estimation for the traditional two-stage set-up, with composite information specified as follows: (i) for the cluster population U_1 , there is a known total $\sum_{U_1} \mathbf{x}_{(c)i}$, where $\mathbf{x}_{(c)i}$ is a vector value associated with the cluster U_i , for $i \in U_1$; (ii) for the population of elements $U = \bigcup_{i \in U_1} U_i$, there is a known total $\sum_U \mathbf{x}_k$, where the vector value \mathbf{x}_k is associated with the element $k \in U$. Suppose both cluster statistics and element statistics are to be produced in the survey: Both the cluster population total $Y_1 = \sum_{U_1} y_{(c)i}$ and the element population total $Y = \sum_U y_k$ are to be estimated.

If no relation is imposed between cluster weights w_{li} and element weights w_k , the former are calibrated to satisfy $\sum_{s_1} w_{li} \mathbf{x}_{(c)i} = \sum_{U_1} \mathbf{x}_{(c)i}$, the latter to satisfy $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. (Here, s_1 is the sample of clusters from U_1 ; s_i is the sample of elements from the cluster U_i ; $s = \bigcup_{i \in s_1} s_i$ is the entire sample of elements.) Then $\hat{Y}_{\text{CAL}} = \sum_{s_1} w_{li} y_{(c)i}$ estimates the cluster population total Y_1 , and $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ estimates the element population total Y .

Integrated weighting is often used in practice: A convenient relationship is imposed between the cluster weight w_{li} and the weights w_k for the elements within the selected cluster. Two forms of integrated weighting are discussed in Estevao and Särndal (2006).

One of these is to impose $w_k = d_{k|l} w_{li}$, where $d_{k|l}$ is the inverse of the probability of selecting element k within cluster i . (For example, in single stage cluster sampling, when all elements k in a sampled cluster are selected, then $d_{k|l} = 1$. Consequently $w_k = w_{li}$ is imposed, and all elements in the cluster receive the same weight for computing element statistics, and that same weight is also used for computing cluster statistics.) The calibration equation $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$ then reads $\sum_{s_1} w_{li} \sum_{s_i} d_{k|l} \mathbf{x}_k = \sum_U \mathbf{x}_k$. The cluster weights w_{li} are now derived by minimizing $\sum_{s_1} (w_{li} - d_{li})^2 / d_{li}$ subject to the calibration equation that takes both kinds of information into account:

$$\left(\begin{array}{c} \sum_{s_1} w_{li} \mathbf{x}_{(c)i} \\ \sum_{s_1} w_{li} \sum_{s_i} d_{k|l} \mathbf{x}_k \end{array} \right) = \left(\begin{array}{c} \sum_{U_1} \mathbf{x}_{(c)i} \\ \sum_U \mathbf{x}_k \end{array} \right). \quad (8.1)$$

Once the w_{li} are determined, the element weights $w_k = d_{k|l} w_{li}$ follow.

Another reasonable integrated weighting is to impose $\sum_{s_i} w_k = N_i w_{li}$. For example, for single stage cluster sampling it implies that the cluster weight w_{li} is the average of the element weights w_k in that cluster.

Two-stage sampling is also the topic in Kim, Breidt and Opsomer (2005). They assume auxiliary information for clusters, via a single quantitative cluster variable $x_{(c)i}$, but none for elements. They develop and examine a GREG type estimator of the element total $Y = \sum_U y_k$, $\hat{Y} = \sum_{i \in U_1} \hat{\mu}_i + \sum_{i \in s_1} d_i (\hat{t}_i - \hat{\mu}_i)$, where \hat{t}_i is design unbiased for the cluster total $t_i = \sum_{U_i} y_k$, and $\hat{\mu}_i$ is obtained by local polynomial regression fit. The estimator can be expressed on the linearly weighted form, with weights that turn out to be calibrated to the population totals of powers of the cluster variable $x_{(c)i}$.

8.3 Household weighting and person weighting

Some important social surveys set the objective to produce both household estimates and person estimates; some study variables are household (cluster) variables, others are person (element) variables. Consequently, a number of papers have addressed the situation with single stage cluster sampling ($d_{k|l} = 1$) and the integrated weighting that gives all members of a selected household equal weight, a weight also used for producing household statistics. A general solution for this weighting problem, when both household information and person information are specified, is to obtain the household weights w_{li} calibrated as in equation (8.1) with $d_{k|l} = 1$, then take $w_k = w_{li}$.

Several articles focus on auxiliary vector values \mathbf{x}_k attributed to persons. Alexander (1987) derives weights by minimizing chi-square distance, whereas Lemaître and Dufour (1987) and Niewenbroek (1993) derive the integrated weights via a GREG estimator. The Lemaître and Dufour technique proceeds by an indirect construction of an "equal shares auxiliary vector value" for all persons in a selected household; their result is derivable from the direct procedure in Section 8.2.

The household-weighting/person-weighting question is revisited in more recent papers. Some authors display calibration thinking, others GREG thinking. Isaki, Tsay and Fuller (2004) formulate the problem as one of calibrated weighting; their weights respect both household controls and person controls; no explicit assisting models are formulated. By contrast, Steel and Clark (2007) proceed by the GREG approach, with linear assisting model statements and accompanying variance structures.

9. Calibration for nonresponse adjustment

9.1 Traditional adjustment for nonresponse

The context of many good theory articles is the simple one of Section 2, which includes total absence of nonresponse. It is good theory for conditions that seldom or never occur. (As an author of papers in that stream, I am not without guilt.) Practically all surveys encounter non-response; although undesirable, it is a natural feature, and theory should incorporate it, from the outset, via a perspective of selection in two phases.

In many surveys, nonresponse rates are extremely high today, compared with what they were 40 years ago, that is, so low that one could essentially ignore the problem. Today, survey sampling theory needs more and more to address the damaging consequences of nonresponse. In particular, one pressing objective is to examine the bias and to try to reduce it as far as possible.

A probability sample s is drawn from $U = \{1, 2, \dots, k, \dots, N\}$; the known design weight of element k is $d_k = 1/\pi_k$. Nonresponse occurs, leaving a response set r , a subset of s ; the study variable value y_k is observed for $k \in r$ only. The unknown response probability of element k is $\Pr(k \in r|s) = \theta_k$. The unbiased estimator $\hat{Y} = \sum_r d_k \phi_k y_k$ is ruled out because $\phi_k = 1/\theta_k$ is unknown. To keep the idea of a linearly weighted sum, how do we then construct the weights? Unit nonresponse adjustment by weighting, based on “nonresponse modeling”, has a long history. Calibration offers a newer perspective.

In what we may call “the traditional procedure”, the probability design weights $d_k = 1/\pi_k$ are first adjusted for nonresponse and possibly for other imperfections such as outliers. The information used for this step is often a grouping of the sampled elements. Finally, if reliable population totals are accessible, the adjusted design weights are subjected to a calibration with respect to those totals.

The methodology of the Labour Force Survey of Canada, described in Statistics Canada (1998), exemplifies this widespread practice. A (modified) design weight is first computed for a given household, as the product of three factors. The product of the design weight and a nonresponse adjustment factor is called the sub-weight. The sub-weights are subjected in the final step to a calibration with respect to postcensal, highly accurate estimates of population by age group, sex and sub-provincial regions. The final weights meet the desirable objective of consistency, in regions within a province, with the postcensal estimates. The nonresponse bias remaining in the resulting estimates is unknown but believed to be modest.

The traditional procedure is embodied in the estimator type $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$, where θ_k has been estimated by $\hat{\theta}_k$ in a preliminary step, using response (propensity)

modeling. What theory demands of the statistician is not an easy task, namely, to formulate “the true response model”, capable of providing accurate, non-biasing values $\hat{\theta}_k$. But the factors $1/\hat{\theta}_k$ are applied in many surveys in an uncritical and mechanical fashion, for example, by straight expansion within the strata already used for sample selection.

The traditional procedure is apparent for example in Ekholm and Laaksonen (1991) and in Rizzo, Kalton and Brick (1996).

Practitioners often act as if the resulting $\hat{Y} = \sum_r d_k (1/\hat{\theta}_k) y_k$ (following a more or less probing response modeling trying to get the $\hat{\theta}_k$) is essentially unbiased, something which it is not (unless the ideal model happens to be specified); one acts (for purposes of variance estimation, for example) as if $\pi_k \hat{\theta}_k$ is the true selection probability of element k in a single step of selection, something which it is definitely not. This practice, with roots in the idyllic past, becomes more and more vulnerable as nonresponse rates continue their surreptitious climb.

An unavoidable bias results from the replacement of θ_k by $\hat{\theta}_k$. Decades ago, when the typical nonresponse was but a few per cent, it was defensible to ignore this bias, but with today's galloping nonresponse rates, the practice becomes untenable. By first principles, unbiased estimation is the goal, not an estimation where the squared bias is a dominating (and unknown) contributor to the Mean Squared Error. We must resolve to limit the bias as much as possible. Calibration reasoning can help in constructing an auxiliary vector that meets this objective.

9.2 Calibration for nonresponse bias adjustment

More or less contrasting with the traditional procedure are a number of recent papers that emphasize calibration reasoning to achieve the nonresponse adjustment. Recent references are Deville (1998, 2002), Ardilly (2006), chapter 3, Skinner (1998), Folsom and Singh (2000), Fuller (2002), Lundström and Särndal (1999), Särndal and Lundström (2005) and Kott (2006).

Calibration reasoning starts by assessing the total available auxiliary information: information at the sample level (auxiliary variable values observed for respondents and for nonrespondents), information at the population level (known population auxiliary totals). The objective is to make the best of the two sources combined, so as to reduce both bias and variance. The design weights are modified, in one or two calibration steps, to make them reflect (i) the outcome of the response phase, (ii) the individual characteristics of the respondents, and (iii) the specified auxiliary information. The information can be summarized as follows:

Population level: The vector value \mathbf{x}_k^* is known (specified in the frame) for every $k \in U$, thus known for every $k \in s$ and for every $k \in r$; $\sum_U \mathbf{x}_k^*$ is a known population total.

Sample level: The vector value \mathbf{x}_k° is known (observed) for every $k \in s$, and thereby known for every $k \in r$; the unknown total $\sum_U \mathbf{x}_k^\circ$ is estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$.

Calibration on this composite information can be done in two steps (intermediate weights computed first, then used in the second step to produce final weights) or directly in one single step. Modest differences only are expected in bias and variance of the estimates. In the single step option, the combined auxiliary vector and the corresponding information are

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}.$$

Using an extension of the instrument vector method in Section 4.3, we seek calibrated weights $w_k = d_k v_k$, where $v_k = F(\lambda' \mathbf{z}_k)$ is the nonresponse adjustment factor, with a vector λ determined through the calibration equation $\sum_r w_k \mathbf{x}_k = \mathbf{X}$; the resulting calibration estimator is $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$. It is enough to specify the instrument vector value \mathbf{z}_k for respondents only; \mathbf{z}_k is allowed to differ from \mathbf{x}_k . The function $F(\cdot)$ has the same role as in Sections 4.2 and 4.3. Here, $F(\lambda' \mathbf{z}_k)$ implicitly estimates the inverse response probability, $\phi_k = 1/\theta_k$ as Deville (2002), Dupont (1995), Kott (2006) have noted. In the linear case, $F(u) = 1 + u$, and $v_k = 1 + \lambda' \mathbf{z}_k$, with $\lambda' = (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k) / (\sum_r d_k \mathbf{z}_k \mathbf{x}_k')^{-1}$.

The variables that make up the vector \mathbf{x}_k° , although observed for sampled elements only, can be crucially important for the reduction of nonresponse bias (although less important than the \mathbf{x}_k^* for the reduction of variance). For example, Beaumont (2005b) discusses data collection process variables can be used in building the \mathbf{x}_k° vector component.

9.3 Building the auxiliary vector

In some surveys, there are many potential auxiliary variables, as pointed out for example by Rizzo, Kalton and Brick (1996), and Särndal and Lundström (2005). For example, for surveys on households and individuals in Scandinavia, a supply of potential auxiliary variables can be derived from a matching of existing high quality administrative registers. A decision then has to be made which of these variables should be selected for inclusion in the auxiliary vector \mathbf{x}_k to make it as effective as possible, for bias reduction in particular. As Rizzo, Kalton and Brick (1996) point out, “the choice of auxiliary variables is ...

probably more important than the choice of the weighting methodology.”

Let us examine the bias, when $\mathbf{z}_k = \mathbf{x}_k$. We need to compare alternative \mathbf{x}_k -vectors in order to finally settle one likely to yield the smallest bias. (I assume \mathbf{x}_k to be such that $\mu' \mathbf{x}_k = 1$ for all k and some constant vector μ , as is the case for many \mathbf{x}_k -vectors, including the examples 1 to 5 at the beginning of Section 2.) A close approximation to the bias of \hat{Y}_{CAL} is obtained by Taylor linearization as *nearbias*(\hat{Y}_{CAL}) = $(\sum_U \mathbf{x}_k)' (\mathbf{B}_{U;0} - \mathbf{B}_U)$, which involves a difference between the weighted regression coefficient $\mathbf{B}_{U;0} = (\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_U \theta_k \mathbf{x}_k y_k$ and the unweighted one, $\mathbf{B}_U = (\sum_U \mathbf{x}_k \mathbf{x}_k')^{-1} (\sum_U \mathbf{x}_k y_k)$. Unless all θ_k are equal, the bias caused by the difference in the two regression vectors may be substantial, even though \mathbf{x}_k is a seemingly “good auxiliary vector”. This expression for *nearbias* is given in Särndal and Lundström (2005); related bias expressions, under different conditions, are found in Bethlehem (1988) and Fuller *et al.* (1994). We can write alternatively *nearbias*(\hat{Y}_{CAL}) = $\sum_U (\theta_k M_k - 1) y_k$, where $M_k = (\sum_U \mathbf{x}_k)' (\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. In comparing possible alternatives \mathbf{x}_k , a convenient benchmark is the “primitive auxiliary vector”, $\mathbf{x}_k = 1$ for all $k \in U$, which gives $\hat{Y}_{\text{CAL}} = N \bar{y}_r = N \sum_r y_k / n_r$, where n_r is the number or respondents, with *nearbias*($N \bar{y}_r$) = $N(\bar{y}_{U;0} - \bar{y}_U)$, where $\bar{y}_{U;0} = \sum_U \theta_k y_k / \sum_U \theta_k$ and $\bar{y}_U = \sum_U y_k / N$. The ratio

$$\text{relbias}(\hat{Y}_{\text{CAL}}) = \frac{\text{nearbias}(\hat{Y}_{\text{CAL}})}{\text{nearbias}(N \bar{y}_r)} = \frac{\sum_U (\theta_k M_k - 1) y_k}{N(\bar{y}_{U;0} - \bar{y}_U)}$$

measures how well a candidate vector \mathbf{x}_k succeeds in controlling the bias, when compared with the primitive vector. We seek an \mathbf{x}_k that will give a small bias. But *relbias*(\hat{Y}_{CAL}) is not a computable bias indicator; it depends on unobserved y_k and on unobservable θ_k . We need a computable indicator that approximates *relbias*(\hat{Y}_{CAL}) and depends on the \mathbf{x} -vector but not on the y -variables, of which the survey may have many.

It is easy to see that *relbias*(\hat{Y}_{CAL}) = 0 if an ideal (probably non-existent) \mathbf{x} -vector could be constructed such that $\phi_k = 1/\theta_k = \lambda' \mathbf{x}_k$ for all $k \in U$ and some constant vector λ .

For an \mathbf{x} -vector that can actually be formed in the survey, we can at least obtain predictions of the ϕ_k : Determine $\hat{\lambda}$ to minimize $\sum_U \theta_k (\phi_k - \hat{\lambda}' \mathbf{x}_k)^2$; we find $\hat{\lambda} = \hat{\lambda}_U$, where $\hat{\lambda}_U' = (\sum_U \mathbf{x}_k) (\sum_U \theta_k \mathbf{x}_k \mathbf{x}_k')^{-1}$; the predicted value of ϕ_k is $\hat{\phi}_{kU} = \hat{\lambda}_U' \mathbf{x}_k = M_k$. The (theta-weighted) first and second moment of the predictions $\hat{\phi}_{kU} = M_k$ are, respectively, $\bar{M}_{U;0} = \sum_U \theta_k M_k / \sum_U \theta_k = N / \sum_U \theta_k = 1/\bar{\theta}_U$ and

$$Q = \frac{1}{\sum_U \theta_k} \sum_U \theta_k (M_k - \bar{M}_{U;0})^2 = (1/\bar{\theta}_U) (\bar{M}_U - 1/\bar{\theta}_U)$$

where $\bar{M}_U = \sum_U M_k / N$. Särndal and Lundström (2007) show that $relbias(\hat{Y}_{CAL})$ and Q have under certain conditions an approximately linear relationship,

$$relbias(\hat{Y}_{CAL}) \approx 1 - \frac{Q}{Q_0}$$

where $\bar{\phi}_U = \sum_U \phi_k / N$ and $Q_0 = (1/\bar{\theta}_U)(\bar{\phi}_U - 1/\bar{\theta}_U)$ is the maximum value of Q . Thus if Q were computable, it could serve as an indicator for comparing the different candidate \mathbf{x}_k -vectors. A computable analogue \hat{Q} of Q is instead obtained as the variance of the corresponding sample-based predictions

$$\hat{\phi}_{ks} = \hat{\lambda}'_s \mathbf{x}_k = (\sum_s d_k \mathbf{x}_k)'$$

$$(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k = m_k, \text{ so that}$$

$$\hat{Q} = \frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r,d})^2 = \bar{m}_{r,d} (\bar{m}_{s,d} - \bar{m}_{r,d})$$

where

$$\bar{m}_{r,d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k}; \bar{m}_{s,d} = \frac{\sum_s d_k m_k}{\sum_s d_k}.$$

We expect $relbias$ to decrease in a roughly linear fashion as \hat{Q} increases; thus, independently of the y -variables, \hat{Q} may be used as a tool for ranking different \mathbf{x} -vectors in regard to their capacity of reduce the bias.

We can use \hat{Q} as a tool to select x -variables for inclusion in the \mathbf{x}_k -vector, for example, by stepwise forward selection, so that variables are added to \mathbf{x}_k one at a time, the variable to enter in a given step being the one that gives the largest increment in \hat{Q} . The method is described in Särndal and Lundström (2007).

10. Calibration to account for other non-sampling error

Nonresponse errors are critical determinants of the quality of published statistics. When we examine how the calibration approach may intervene in the treatment other sources of non-sampling error than the nonresponse, the literature to date is not surprisingly much less extensive. However, several authors sketch a calibration reasoning to also incorporate frame errors, measurement errors, and outliers. Calibration has a potential to provide a more general theory for estimation in surveys, encompassing the various non-sampling errors.

As Deville (2004) points out (my translation from the French): "The concept of calibration lends itself to be applied with ease and efficiency to a great variety of problems in survey sampling. Its scope goes beyond that of regression estimation, an idea to which some seem to wish to reduce the calibration approach". He provides a brief

sketch of how a treatment of several of the nonresponse errors may be accomplished under the caption of calibration thinking.

Folsom and Singh (2000) present a weight calibration method using what they call the generalized exponential model (GEM). It deals with three aspects: extreme value treatment, nonresponse adjustment and calibration through post-stratification. The method provides built-in control for extreme values. Calibration to treat both coverage errors (under- or over-coverage of the frame) and nonresponse is discussed in Särndal and Lundström (2005) and Kott (2006). Skinner (1998) discusses uses of calibration in the presence of nonresponse and measurement error. He notes something which remains a challenge almost ten years later: "More research is needed to investigate the properties of calibration estimates in the presence of non-sampling errors".

11. Conclusion

If I am to select one issue for a concluding reflection on the contents of this paper, let me focus on the concept of auxiliary information. It is the pivotal concept in the paper. If there is not auxiliary information, there is no calibration approach; there is nothing to calibrate on. I noted on the other hand that regression (GREG) estimation is an alternative but different thought process for putting auxiliary information to work in the estimation.

An objective in this paper has been to give a portrait of the two types of reasoning, and I made a point of noting how the thinking differs. I gave examples where essentially the same estimation objective is tackled by some authors through calibration reasoning, by others through GREG reasoning (or at least *primarily* by one or the other type). The respective estimators that they end up recommending may or may not agree. Whether or not the difference has significant consequence (for variance, for bias, for practical matters such as consistency and transparency) depends on the situation. This paper may help contributing an awareness of the separation existing between two thought processes that have guided researchers survey sampling.

References

- Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- Andersson, P.G., and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31, 95-99.
- Ardilley, P. (2006). *Les techniques de sondage*. Paris: Editions Technip.

- Bankier, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Working paper, Census Operations Section, Social Surveys Methods Division, Statistics Canada.
- Bankier, M., Houle, A.M. and Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings, Section on Survey Research Methods*, American Statistical Association, 66-75.
- Beaumont, J.-F. (2005a). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.
- Beaumont, J.-F. (2005b). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31, 227-231.
- Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195-208.
- Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.
- Bethlehem, J.G., and Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.
- Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.
- Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.
- Deming, W.E. (1943). *Statistical Adjustment of Data*. New York: John Wiley & Sons, Inc.
- Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30, 17-26.
- Deville, J.C. (1998). La correction de la nonréponse par calage ou par échantillonnage équilibré. Paper presented at the Congrès de l'ACFAS, Sherbrooke, Québec.
- Deville, J.C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.
- Deville, J.C. (2004). Calage, calage généralisé et hypercalage. Internal document, INSEE, Paris.
- Deville, J.C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J.C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, 25, 43-56.
- Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-135.
- Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.
- Estevao, V.M., and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.
- Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.
- Estevao, V.M., and Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20, 645-660.
- Estevao, V.M., and Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74, 127-147.
- Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, 60, 3-21.
- Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. *Proceedings, Section on Survey Research Methods*, American Statistical Association, 598-603.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 45-51.
- Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Harms, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. Chintex working paper no. 13, Federal Statistical Office, Germany.
- Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.
- Hidioglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.
- Hidioglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Huang, E.T., and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. *Proceedings, Social Statistics Section*, American Statistical Association, 300-305.
- Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2004). Weighting sample data subject to independent controls. *Survey Methodology*, 30, 35-44.
- Kalton, G., and Flores-Cervantes, I. (1998). Weighting methods. In *New Methods for Survey Research* (Eds. A. Westlake, J. Martin, M. Rigg and C. Skinner), Berkeley, U.K.: Association for Survey Computing.
- Kim, J., Breidt, F.J. and Opsomer, J.D. (2005). Nonparametric regression estimation of finite population totals under two-stage sampling. Unpublished manuscript.
- Knottnerus, P., and van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565-584.

- Kott, P.S. (2004). Comment on Demnati and Rao: Linearization variance estimators for survey data. *Survey Methodology*, 30, 27-28.
- Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.
- Kovačević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 139-144.
- Krapavickaitė, D., and Plikusas, A. (2005). Estimation of a ratio in the finite population. *Informatica*, 16, 347-364.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.
- LeGuennec, J., and Sautory, O. (2002). CALMAR2: une nouvelle version de la macro CALMAR de redressement d'échantillon par calage. *Actes des Journées de Méthodologie*, INSEE, Paris.
- Lehtonen, R., and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51-55.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44.
- Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.
- Lemaître, G., and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.
- Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.
- Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.
- Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.
- Montanari, G.E., and Ranalli, M.G. (2003). On calibration methods for design-based finite population inferences. Bulletin of the International Statistical Institute, 54th session, volume LX, contributed papers, book 2, 81-82.
- Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model-calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.
- Myrskylä, M. (2007). Generalised regression estimation for domain class frequencies. Statistics Finland Research Reports 247.
- Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Report, Central Bureau of Statistics, The Netherlands.
- Nieuwenbroek, N.J., and Boonstra, H.J. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. The Survey Statistician, Software Reviews, July 2002.
- Nieuwenbroek, N.J., Renssen, R.H. and Hofman, L. (2000). Towards a generalized weighting system. In *Proceedings, Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.
- Park, M. and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. *Survey Methodology*, 31, 85-93.
- Plikusas, A. (2006). Non-linear calibration. Proceedings, Workshop on Survey Sampling, Ventspils, Latvia. Riga: Central Statistical Bureau of Latvia.
- Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.
- Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.
- Renssen, R.H., Kroese, A.H. and Willeboordse, A.J. (2001). Aligning estimates by repeated weighting. Report, Central Bureau of Statistics, The Netherlands.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.
- Rueda, M., Martínez, S., Martínez, H. and Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.
- Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model-assisted Survey Sampling*. New York: Springer-Verlag.
- Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.
- Särndal, C.-E., and Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. Statistics Sweden: Research and development - methodology report 2007:2, to appear, *Journal of Official Statistics*.
- Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.
- Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, 24, 41-50.
- Skinner, C. (1998). Calibration weighting and non-sampling errors. *Proceedings International Seminar on New Techniques for Statistics*, Sorrento, November 4-6, 1998, 55-62.
- Statistics Canada (1998). Methodology of the Canadian Labour Force Survey. Statistics Canada, Household Survey Methods Division. Ottawa: Minister of Industry, catalogue no. 71-526-XPB.
- Statistics Canada (2003). Quality Guidelines (fourth edition). Ottawa: Minister of Industry, Catalogue no. 12-539-XIE.
- Steel, D.G., and Clark, R.G. (2007). Person-level and household-level regression estimation in household surveys. *Survey Methodology*, 33, 51-60.
- Stukel, D.M., Hidirolou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.
- Théberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.
- Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26, 99-107.

- Tillé, Y. (2002). Unbiased estimation by calibration on distribution in simple sampling designs without replacement. *Survey Methodology*, 28, 77-85.
- Tracy, D.S., Singh, S. and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, 29, 99-104.
- Vanderhoeft, C. (2001). Generalised calibration at Statistics Belgium. SPSS Module g-CALIB-S and current practices. Statistics Belgium Working Paper no. 3. Available at: www.statbel.fgov.be/studies/paper03_en.asp.
- Vanderhoeft, C., Waeytens, E. and Museux, J.M. (2001). Generalised calibration with SPSS 9.0 for Windows baser. In *Enquêtes, Modèles et Applications* (Eds. J.J. Droesbeke and L. Lebart), Paris: Dunod
- Webber, M., Latouche, M. and Rancourt, E. (2000). Harmonised calibration of income statistics. Statistics Canada, internal document, April 2000.
- Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Zheng, H., and Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Zieschang, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

Weighting for two-phase surveyed data

Seppo Laaksonen¹

Abstract

Missingness may occur in various forms. In this paper, we consider unit non-response, and hence make attempts for adjustments by appropriate weighting. Our empirical case concerns two-phase sampling so that first, a large sample survey was conducted using a fairly general questionnaire. At the end of this contact the interviewer asked whether the respondent was willing to participate in the second phase survey with a more detailed questionnaire concentrating on some themes of the first survey. This procedure leads to three missingness mechanisms. Our problem is how to weight the second survey respondents as correctly as possible so that the results from this survey are consistent with those obtained with the first phase survey. The paper first analyses missingness differences in these three steps using a human survey dataset, and then compares different weighting approaches. Our recommendation is that all available auxiliary data should have been used in the best way. This works well with a mixture of the two classic methods that first exploits response propensity weighting and then calibrates these weights to the known population distributions.

Key Words: Calibration; Internal vs. external auxiliary variables; Response propensity modelling method; Selective sub-sample.

1. Introduction

A standard survey is composed of one step or phase. This means that the potential survey units have first been chosen using a certain sampling design, and attempts have then been made to contact and interview these units as well as possible. However, varying amounts of non-response or other forms of missingness or data deficiencies will have occurred. Usually, addressing missingness leads to the application of post-survey adjustment methods of varying degrees of sophistication, which take advantage of available auxiliary variables. The auxiliary variables may be derived from various sources (see *e.g.*, Laaksonen 1999 or an extended version in Laaksonen 2006b, and Lundström and Särndal 2001), but for weighting purposes these are usually taken from registers, or other administrative sources or surveys. These kinds of auxiliary variables could be called *external*, if we want to distinguish them from *internal* auxiliary variables, that is, internal in the sense that the information is derived from the same survey or from its predecessor.

Internal auxiliary variables are especially used for imputations when the values for some items are missing. Such variables are also extensively used in panel surveys if a certain respondent has responded in one wave but not in another. In panel surveys, internal auxiliary information may be used both for weighting adjustments and for imputations.

This paper does not concern a standard survey as described above. It discusses two special characteristics:

- (i) A survey consisting of two (or, in some sense, three) steps or phases. The first phase is like a standard survey, in which a certain number of units respond. For the second phase, we only keep in the frame the respondents who are willing to contribute to a more detailed survey. This leads first to having to distinguish such first-phase respondents who say they are willing to participate voluntarily, on one hand, and those of these respondents who actually answer the second questionnaire. This being the case, the latter subgroup will thus respond to both questionnaires.
- (ii) When making attempts for post-survey adjustments, we will have the option to exploit both external and internal auxiliary variables in the second phase. The internal variables will thus be available from the first survey.

We are only considering weighting adjustments although some of our ideas could be used in imputations, too. The approach of the paper has not been much used in cross-sectional surveys although the same problem has been often met. For instance, it is typical that a face-to-face survey is conducted first and that at the end of it the interviewer will request the interviewee to respond to a self-administered additional questionnaire and, if the respondent is willing, the interviewer will hand out the questionnaire immediately for filling in, or submit it later to the volunteer. In both cases, the answers will be received by post or email. A recent example of this type is the European Social Survey (ESS) in which the supplementary questionnaire concerns especially the values of life (see www.Europeansocialsurvey.com).

1. Seppo Laaksonen, Department of Mathematics and Statistics, Box 68, FIN-00014 University of Helsinki, Finland. E-mail: Seppo.Laaksonen@Helsinki.Fi.

Naturally, not all face-to-face respondents fill in this questionnaire.

The second-phase questions do not necessarily concern the same topic as those of the core questionnaire. Another usual strategy is to start with a broad questionnaire on a specific subject and then continue in the second phase with more detailed questions about the same subject. There can be some feedback from the first phase to the second questionnaire, and even to a sample that depends on the distribution of the key variables of the first survey (this is an example of adaptive sampling). This is often the case where there is not much experience in this type of a survey. Thus, the first survey also plays the role of a pilot survey. The so-called master samples are also close to this idea whereby the first phase survey (including a sample from administrative sources, like a micro census in some countries) may be conducted for constructing an appropriate sampling frame. In this case, the variables of the master sample are fairly limited, including usually only factual or background information.

In the case of master sampling, the objective is that the constructed sampling frame is a good representation of the target population. Hence, when going to a sample, this frame information can be well used as auxiliary data for editing, imputing and weighting of the real survey (second phase survey). Each real survey is thus a sub-sample of the master sample. We consider here a more complex case as illustrated by Table 1.

Table 1 Illustration of initial sample and three follow-up datasets

Sample with Auxiliary Variables X (<i>Gender, Age, Region, Season</i>)	First Phase Respondents with Variables Y_1 (e.g., <i>Health, Outdoor</i>)	Volunteers for the Second Phase with Variables Y_1	Second Phase Respondents with Variables Y_2 (e.g., <i>Skating, Boating</i>)
Design Weights for 12,554 Units without Overcoverage	Basic Weights for 10,666 Units	Weights for 8,481 Units	Weights for 5,480 Units

First, there is a standard sampling procedure including some auxiliary variables X . A fairly high response rate was obtained (10,666 out of 12,554 units, about 85%) for the first survey. Some attrition occurred due to the fact that all respondents were not willing to participate in the second survey (we now have 68% of the initial sample left). Due to a rather high non-response rate in this second phase (in spite of voluntariness), our remaining sub-sample covers only 44

per cent of the initial sample. We now have the following three datasets available for the analysis:

- A. First-phase respondents with survey variables Y_1
- B. Second-phase respondents with survey variables Y_2
- C. Both first and second-phase respondents with survey variables Y_1 and Y_2 .

Most users will receive both files, A and B , and they can merge these together and obtain file C if a common identifier is available. What does a user expect having received both data files? Naturally, that the estimate for the same parameter from both files is as identical as possible, that is, the results are consistent with each other. The user obviously understands that a certain parameter estimated from the smaller file C is less accurate than that estimated from a larger file. In principle, it is possible to impute the missing values for variables Y_2 , but we do not believe that it is possible to do this well, hence we approach this question by weighting. Our aim is to attempt to construct adjusted sampling weights for file B so that the analysis over variables Y_1 and Y_2 is as adequate as possible.

Several strategies can be used for this kind of weighting. Useful general aspects have been presented by, among others, Kalton and Kasprzyk (1986), Little (1986), Särndal, Swensson and Wretman (1992), Fuller, Loughin and Baker (1994), Wu and Sitter (2001), and Lundström and Särndal (2001). If we assume that the missingness only depends on the sampling design, we can construct the weights for files A and B in the respective way. For example, if stratified random sampling has been applied, the same stratification would naturally be applied to both phases. In the case of post-stratification, an analogous strategy may be applied.

In our particular example survey, the sampling frame contained the respondents of the first rotation group of the 12 months of the Finnish LFS. Each monthly sample was drawn randomly. The LFS is based on simple random sampling, but due to nonresponse these weights were adjusted by a standard calibration technique (Deville, Särndal and Sautory 1993) using *gender*, *age group* (six categories) and *region* (five categories) as auxiliary variables. Later, we refer to these as design weights. The basic weights for the first-phase respondents were constructed correspondingly adding variable *season* ($4 \times 2 = 8$ categories over two years) to the pattern of auxiliary variables. The ‘*season*’ variable is rarely used in Finnish surveys but was here considered to be necessary due to the ‘seasonal’ nature of the survey (see Section 2). The present paper does not consider this aspect in detail. The first three variables are usual in Finnish human surveys because such information can be validated well from updates in the population register. This being the case, we now presume

that we will have the best possible estimates for the first-phase respondents when using these adjusted calibrated weights. In any case, we have no further access to other possible useful information from external sources.

It is possible to use estimates obtained from the first-phase respondents as benchmarking information to calibrate the second-phase weights. This strategy is not difficult as such, but all variables X and as many of the variables Y_1 as possible should be included in this process. Moreover, as precise aggregate or domain levels as possible should be found for this strategy, which is not an easy job and hence not attempted in this study. Nevertheless, it is not guaranteed that the estimates for other aggregates will be unbiased enough (results in Laaksonen 1999 give some evidence for this conclusion).

My proposed strategy is more straightforward and works without technical problems for all domains although it is not of course guaranteed that a possible bias will be substantially reduced for all domains. Thus, I have not tried any advanced calibration strategy, although this could be workable. I hope that other authors will show its possible benefits. A useful reference for them is the paper by Dupont (1995) that considers calibration of two-phase survey data, however without empirical evidence. It should be noted that I use calibration, but not a very advanced one (see Section 3).

The proposed methodology of this paper is largely based on a response propensity modelling that has been successfully used in other types of situations, see e.g., Ekholm and Laaksonen (1991), Laaksonen (1999), Duncan and Stasny (2001), and Laaksonen and Chambers (2006). The situation of Rizzo, Kalton and Brick (1996) is fairly close to the two-phase case of this paper, although it is concerned with a panel. Their methodology also has some similar features. In addition, a major difference concerns the response mechanism that here occurs in two steps, that is, both due to voluntariness and due to response in the second phase. We analyse these steps separately, too. Naturally, we compare the results obtained with alternative techniques. In Section 2, we briefly further describe our surveys and datasets, and Section 3 details the principles of our methods. Section 4 presents comparison results, and Section 5 draws a conclusion.

2. Principles of the datasets

The data are from a special survey conducted among Finnish citizens aged 15-74 years old (for more information, see Virtanen, Pouta, Siev  nen and Laaksonen 2001). The topic concerns their leisure time activities, especially relating to outdoor hobbies and activities. First, a CAPI (computer-assisted personal interview) survey was

conducted, covering various leisure time and hobby questions such as cycling, motorcycling, walking, jogging, sailing, swimming, hunting, fishing, nature photography, skiing, skating and riding. In all cases, the reference period was the previous year. Second, at the end of this survey the respondents were asked whether they would be willing to receive a special postal survey questionnaire in which more detailed questions would be asked about some of these activities. This survey would be conducted in a few weeks' time.

The survey was conducted over two years (1998-2000) in order to reduce response and interviewing burden. Another reason for this was that since these activities are seasonal to some extent, the responses were expected to be seasonally influenced (e.g., responses to questions about skiing might be different in summer and winter). The initial sample size after the removal of overcoverage (104 units of overcoverage) was 12,554 individuals.

We chose the following binary variables for our analysis presented in Section 4: *Outdoors* (person has performed regularly some outdoor activities in the nature), *Health* (is health good enough for outdoor activities?), *Skiing*, *Fishing*, *Skating*, *Boating*, *Cycling* and *Jogging*. In all cases, value = 1 means that a person has engaged in the activity during the preceding year, and value = 0, respectively, the opposite. All these variables were included in the first-phase questionnaire and we hence knew what to expect after the two consecutive phases. Note, that there are more complex variables in the data set but this simpler choice was made in order to interpret results more easily. The main conclusions are the same in the case of another choice.

In Section 4, we present two types of comparison, (i) those based on known information from the first phase, and (ii) those not based on known information. In both cases, we can fortunately check how well we have succeeded in the reduction of bias since we actually know the 'true' (or best possible) estimates. In addition, we analyse some variables only included in the second questionnaire, but we cannot say definitely how well each method works in these cases. We do not present the latest results in detail but these were observed to behave similarly to those of the second approach.

3. Response propensity modelling method and calibration

This study comprises three steps with the following weighting specifications:

First, well-designed calibrated sampling weights for the first phase respondents were created using the variables *Region*, *Gender*, *Age group* and *Season* (see also Section 1). Let us use symbol w_k for these sampling weights for

respondent k . These weights thus are based on calibration, and also called ‘Basic’. Note that before this we have, before naturally, constructed *design weights* for the dataset, based on the stratified random sampling design. These are thus available for the non-respondents of the first phase, too.

Next, we model voluntariness/response probabilities using the most common link function (Logit is not necessarily the best link function as learned from (2006a), but this is what we use here.), that is, $\text{logit} = \log(\pi/(1 - \pi))$, in which π is the binary response probability (either 1 = volunteer vs. 0 = non-volunteer or 1 = respondent vs. 0 = non-respondent) and the explanatory variables consist of variables X and some variables Y_1 that have been considered to be ‘good.’ The model gives the predicted response probabilities p_k that are now used in the following way when constructing each particular adjusted sampling weight:

$$w_k(\text{res}) = \frac{w_k}{p_k} g_c.$$

Here $g_c = a$ scaling factor which benchmarks the weights to certain known aggregates at level c . There are several alternatives for this benchmarking, but some type of calibration could be considered as a standard way. In this study we use post-stratified aggregates h (this being the cross-classified cell of all three X variables = *Age group*, *Gender* and *Region*, the whole number of cells = $6 \times 2 \times 5 = 60$) using the following straightforward technique

$$g_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}.$$

As already pointed out, the quality is high in Finland for these kinds of post-stratified aggregates but not necessarily for any other aggregates.

Because we have two steps for the second phase, we have the following three model options that were all also used in Section 4:

- (a) Model for voluntariness
- (b) Model for the response given that the person volunteered (called also ‘*TwoStep*’).
- (c) Model for the response as one step (called also ‘*Direct*’ and ‘*OneStep*’).

Note that steps (a) and (b) together give the weights for file B. This leads to the following formulation (vol = volunteer; p_1 = estimated response probability at step 1, p_2 = estimated response probability at step 2; respectively for the scaling factors g_1 and g_2):

$$\text{Step 1: } w_k(\text{vol}) = \frac{w_k}{p_{1k}} g_{1h},$$

$$\text{Step 2: } w_k(\text{res}) = \frac{w_k(\text{vol})}{p_{2k}} g_{2h}.$$

The correct sampling weights had to be used in each modelling task. For models (a) and (c) this meant weights w_k , but for model (b) weights $w_k(\text{vol})$. In our comparison tests we also modelled the first-phase response and here we used design weights. The use of weights in the modelling gives more correct estimates, since we are trying to make our analysis representative for the target population. In some cases, the influence of weights is substantial, like in business surveys where weights often vary more than in standard household surveys. In this case, the results between weighted and unweighted models were not highly different, although the weighted ones should be used (this is well justified in Laaksonen and Chambers 2006 in which the influence of weights is substantial; Rizzo *et al.* 1996 also use weights). The empirical results (estimates, their standard errors and response probabilities) for the weighted solutions are presented later in Section 4.

In addition to our above key techniques, in the next Section we also use weights w_k when providing our ‘*best possible*’ estimates for such parameters that are known, thus based on variables Y_1 .

Moreover, we compare our specific results using post-stratified calibration only without modelling (we use also symbol ‘*cal*’ in the remaining sections). The latter could be interpreted as a very standard way of approaching the weighting problem (this was a house style prior the methodology proposed here). Note, however, that if a response model only includes the variables (and the same categories) used in post-stratification, the response propensity-based weights are exactly the same as obtained by post-stratification.

4. Empirical results

This Section presents results from different methods. First, we give results from different response models and then go on to compare different weights with each other, and at the end of the Section compare some parameter estimates based on different techniques.

4.1 Models for voluntariness and response

In order to fully understand the behaviour of missingness (due to non-response and voluntariness) in all three phases of the survey, we present in Table 2 results that are based on such auxiliary variables X that are available in each step (in practice, we also used the variable ‘season’ but do not include its effects in this analysis since it is not a key issue in this paper).

Table 2

Logistic regressions using the three common explanatory variables in the three phases, that is, for the first phase respondents, for the voluntary respondents in the second phase and for the real respondents in the second phase. The estimates are odds ratios; their 95% confidence intervals are presented in parenthesis

Explanatory variables and other statistics	Model 1 First phase response	Model 2a Voluntariness	Model 3a Response for volunteers	Model 4a Second phase response
<i>Gender</i> (ref. Female)				
Male	0.71 (0.65, 0.78)	0.84 (0.76, 0.93)	0.75 (0.68, 0.83)	0.77 (0.71, 0.83)
<i>Age group</i> (65+)				
24 and under	1.00 (0.83, 1.21)	5.57 (4.65, 6.68)	0.51 (0.41, 0.64)	1.49 (1.28, 1.73)
25-34	0.96 (0.79, 1.15)	4.76 (4.00, 4.81)	0.65 (0.52, 0.81)	1.73 (1.49, 2.00)
35-44	0.85 (0.71, 1.02)	4.08 (3.46, 4.81)	0.64 (0.52, 0.80)	1.62 (1.40, 1.88)
45-54	0.89 (0.74, 1.07)	3.16 (2.71, 3.69)	0.86 (0.69, 1.06)	1.82 (1.58, 2.10)
55-64	1.18 (0.96, 1.45)	2.05 (1.74, 2.41)	1.15 (0.90, 1.47)	1.75 (1.49, 2.04)
<i>Region</i> (North)				
South-East	0.55 (0.46, 0.66)	2.12 (1.79, 2.50)	0.96 (0.79, 1.16)	1.35 (1.17, 1.55)
South-West	0.76 (0.64, 0.91)	1.83 (1.57, 2.14)	1.04 (0.86, 1.25)	1.35 (1.18, 1.55)
Mid-West	1.14 (0.91, 1.42)	2.14 (1.77, 2.59)	1.16 (0.93, 1.43)	1.56 (1.33, 1.83)
Mid-East	0.96 (0.78, 1.18)	1.20 (1.01, 1.44)	1.15 (0.92, 1.43)	1.19 (1.02, 1.40)
Number of observations	12,554	10,666	8,481	10,666
-2 Log L	10,904	10,296	8,569	14,618

There are many interesting outcomes in these consecutive missingness behaviour models. The results of the first survey are fairly ordinary, for example, men respond more poorly than women in both phases. The response propensities are also lower in the South than in the rest of the country. The differences between age groups are somewhat surprising since the middle-age groups respond most poorly.

The voluntariness estimates are different. People in the Mid-East and North are the least willing to participate in the second survey, but the response premia given that a person is voluntary do not differ much. By age, it seems that younger people are more willing to participate but do not, nevertheless, respond very well. Older people, thus, seem in this sense to be more prepared to make a commitment than young people. However, we see clearly that the oldest ones will be under-represented without adjustments.

When considering the two first internal auxiliary variables (Table 3), it is observed that the people who are not relatively healthy (variable *Health*) and who do not

actively pursue recreation in nature (*Outdoor*), are not willing to receive any new questionnaire, either. This is seen from the very high odds ratios. Interestingly, the respective odds ratio for the variable *Health* is close to the one for the volunteers. This, thus, means that a non-healthy person is not very likely to volunteer, but if he/she does, she/he responds as well as a healthy one. The tendency is similar with the variable *Outdoor*. It should be noted that the non-healthy and non-outdoor domains are not very large and although their roles in the response propensity modelling are important, their impacts on the final estimates are not very dramatic (Section 4.3).

When adding the other two internal auxiliary variables, that is, *Skiing* and *Fishing*, the same selectiveness continues although not as substantial. As a conclusion, we see clearly that the response mechanism of the second survey does not seem to be very non-informative. Consequently, it is expected that this leads to some effects on reweights and on survey estimates. These are considered in the next two sub-sections.

Table 3

Logistic regressions using some auxiliary variables from the first phase respondents in addition to those used in Table 2. The model numbers in this table and in Table 2 correspond to each other so that the response variable and the datasets are the same

Explanatory variables and other statistics	Model 2b Voluntariness	Model 3b Response for volunteers	Model 4b Second phase response	Model 4c Second phase response
<i>Gender</i> (ref. Female)				
Male	0.94 (0.85, 1.04)	0.77 (0.69, 0.85)	0.82 (0.75, 0.88)	0.75 (0.68, 0.83)
<i>Age group</i> (65+)				
24 and under	4.92 (4.07, 5.97)	0.52 (0.41, 0.65)	1.30 (1.12, 1.52)	0.51 (0.41, 0.64)
25-34	3.83 (3.18, 4.60)	0.65 (0.52, 0.81)	1.46 (1.25, 1.70)	0.65 (0.52, 0.81)
35-44	3.26 (2.74, 3.88)	0.64 (0.51, 0.80)	1.37 (1.18, 1.58)	0.64 (0.52, 0.80)
45-54	2.59 (2.20, 3.05)	0.85 (0.68, 1.06)	1.56 (1.34, 1.81)	0.86 (0.69, 1.06)
55-64	1.73 (1.45, 2.05)	1.18 (0.89, 1.46)	1.55 (1.32, 1.81)	1.15 (0.90, 1.47)
<i>Region</i> (North)				
South-East	2.15 (1.81, 2.55)	0.96 (0.79, 1.16)	1.34 (1.16, 1.54)	0.96 (0.79, 1.16)
South-West	1.92 (1.64, 2.26)	1.04 (0.86, 1.25)	1.36 (1.19, 1.56)	1.04 (0.86, 1.25)
Mid-West	2.09 (1.71, 2.54)	1.15 (0.93, 1.43)	1.52 (1.29, 1.78)	1.16 (0.93, 1.43)
Mid-East	1.17 (0.98, 1.41)	1.15 (0.91, 1.42)	1.18 (1.00, 1.38)	1.15 (0.92, 1.43)
<i>Outdoor</i>	3.04 (3.43, 2.71)	1.24 (1.43, 1.07)	1.93 (2.15, 1.74)	1.77 (1.97, 1.59)
<i>Health</i>	3.61 (4.61, 2.82)	1.02 (1.58, 0.66)	2.71 (3.51, 2.09)	2.49 (3.24, 1.92)
<i>Skiing</i>				1.36 (1.47, 1.25)
<i>Fishing</i>				1.27 (1.38, 1.17)
Number of observations	10,666	8,481	10,666	10,666
-2 Log L	9,721	8,560	14,342	14,244

4.2 Comparison of different weights

As already explained, we provided several weights. Table 4 gives a summary of these with descriptive statistics in order to explain the changes that occur after each adjustment operation. The design weights cannot be used in our comparisons since no data on variables Y are available for the initial sample. It is, however, illustrative to see that it has the lowest relative variation measured here with $1 + cv^2$ in which cv is the coefficient of variation. This formula is also used as an approximation of the design effect (DEFF). Rizzo *et al.* (1996) also use this indicator when comparing their weights.

The changes are not dramatic in the first step, that is, from design weights to first-phase basic weights (except for the average that is related to decreasing counts), but in the following two steps the DEFFs are higher. We also see that the variation for both calibrated weights is lower than that for the respective response propensity-based weights. The

distribution for each weight is skewed to the right, least for the design weights, naturally. It is somewhat surprising that the skewness is the highest for the volunteer weights. More details about the weight distributions and the differences between the weights are presented in Figures 1 to 3.

Figure 1 illustrates well how some weights have increased substantially due to the response propensity modelling (Model 2b). It is possible to look in detail to see which types of units are under the plots with high weight increase. For example, behind the separate left-side plots with RP weights higher than 700 are persons who are not healthy and do not engage much in outdoor activities but are, nevertheless, still in the volunteer data file. Similarly, we can find other interesting groups by using the results from the model estimations. However, the majority of the plots are in the same area and, consequently, less changes can be expected in the estimates than in the area with more substantial weight changes.

Table 4 Descriptive statistics for different sampling weights. *RP* = Response Propensity

Weight	Phase	Unit size	Average	Skewness	1 + cv ²
Design Weight	Zero	12,658	308	0.94	1.30
(Calibrated) Basic Weight	First	10,666	365	1.30	1.39
Calibrated Weight	Volunteers	8,481	460	2.52	1.63
RP Weight, Model 2b	Volunteers	8,481	460	4.60	1.82
Calibrated Weight	Second	5,480	712	1.64	1.62
TwoStep RP Weight, Models 2b and 3b	Second	5,480	712	3.60	1.84
OneStep RP Weight, Model 4b	Second	5,480	712	2.56	1.80

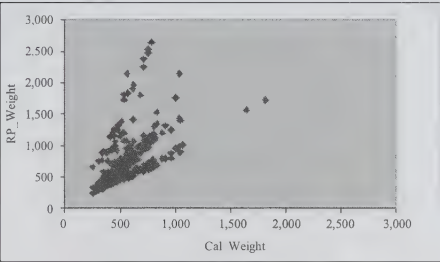


Figure 1 Scatter plot between the two volunteer weights

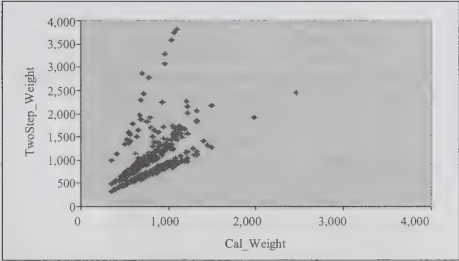


Figure 2 Scatter plot for second phase respondents between the calibrated weight and the two-step response propensity based weight

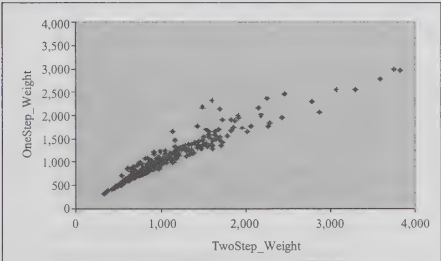


Figure 3 Scatter plot for second phase respondents between the two alternative response propensity modelling weights

The dispersion in the Figure 2 scatter is somewhat stronger than that in Figure 1 but the profile is similar. Consequently, interesting sub-groups can be found behind distinct plots.

Finally, Figure 3 compares the two alternative second phase weights with each other. This scatter differs considerably from the previous two, since the relationship is rather linear. The maximum values of the two-step weights are higher than those of the one-step weights, but the weights of many one-step weights are, however, clearly higher. For example, non-healthy people who, however, engage in outdoor activities receive relatively high one-step weights but there is no clear age effect. On the other hand, people with little outdoor activities in older age groups receive relatively high two-step weights but health does not relate to them. Nevertheless, it is not expected that there will be big differences in respective estimates although one of these two alternatives should have been introduced into use. If this choice were a simpler one, that is, one-step weighting, it would still be useful to analyse both steps and their response propensities separately in order to understand better the reasons for both types of missingness.

4.3 Comparison of parameter estimates

We have not been able to make complete simulation studies with different assumptions in order to analyse which type of method would be best in each particular case. Fortunately, we can get quite close to this by comparing the effects on the estimates from three different perspectives. First, we have prepared the response/voluntariness models by using both X and some Y_1 variables. Consequently, we know the ‘best’ parameter estimates based on these Y_1 values from the first survey. Second, we add auxiliary variables Y_1 in the model but exclude some Y_1 values from them. However, we know the ‘best’ values in these cases and thus can make exact comparisons. Third, we can compare some estimates that are not known in any way. In

this last case, we can only deduce which values might be the best.

We present our explicit results based on the variables described in Section 2. Note that we do not consider it important to present standard errors for each estimate because we are concentrating on the biases in these estimates. However, it is good to notice that the standard errors are around 0.2-0.4 percentage points for the first-phase data set and around 0.3-0.5 percentage points for the second-phase data set (lowest always in *Health*, second in *Jogging*, and highest in *Outdoor* and *Skiing*).

Figure 4 gives the results based on the weights without using any adjustment (that is often the case in practice, unfortunately). We see that the bias is substantial in most estimates, lowest in *Jogging*, which was not very actively practised when compared to *Outdoor*, for example. In general, most users are unhappy with such big biases that are statistically significant and highly significant except for *Jogging* in the second phase (e.g., the 95% confidence interval of the bias for *Health* is from 1.7 points to 2.3 points). Here, as in later results, the bias means *over-estimation* so that while missingness increases the estimate becomes too high. The results without good adjustments will be too optimistic, that is, people seem to do 'too much' of all exercises. Note that the same tendency is obviously also in the first-phase estimates but we cannot justify this. There are surprising differences between those two estimates, sometimes the 'volunteer' data give a more biased result, sometimes it takes the second-phase respondents data. We do not interpret these in detail but naturally they reflect differences in missingness, and can be considered to be warnings for a user.

For comparison, we show again in Figure 5 the same unadjusted results for volunteers as in Figure 4 but we have added the corresponding estimates based on post-stratified calibration and response propensity modelling. This graph clearly shows that post-stratification gives some benefit compared to the unadjusted solution. However, the response propensity method is the best in each case, and extremely good for *Health* and *Outdoor* that have been used as auxiliary variables in the supported models.

Figures 6 and 7 concern the final-step estimates and are thus the most important. Figure 6 shows the same conclusion as Figure 5 in the sense that the response propensity technique is superior to post-stratified calibration although all differences are not statistically highly significant (especially *Jogging*). The difference between the one-step method and the two-step method is fairly small and the bias varies from one variable to the next. Hence, basing on this study, we cannot say which of these two specifications is better.

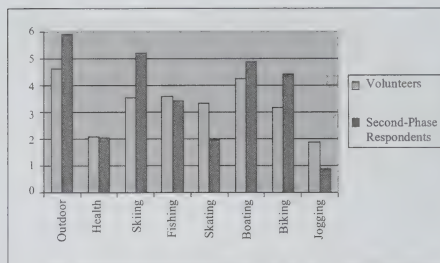


Figure 4 Bias in estimates in percentage points based on unadjusted sampling weights for second-phase respondents and for volunteers

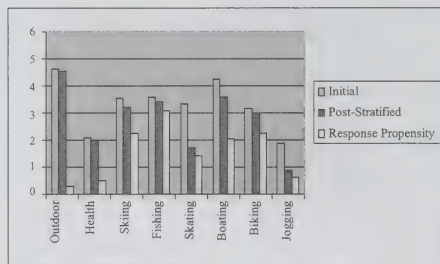


Figure 5 Bias in estimates in percentage points for volunteers based on unadjusted sampling weights (symbol = 'Initial'), post-stratified calibration and response propensity method in which variables *Outdoor* and *Health* have been used as auxiliary variables (Model 2b in Table 3)

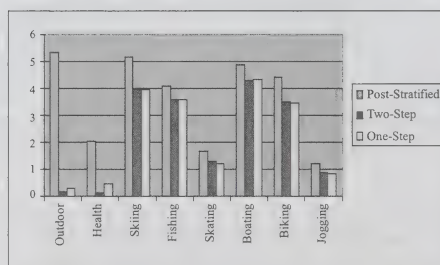


Figure 6 Bias in estimates in percentage points for respondents after both steps based on post-stratification, and the two response propensity methods, i.e., 'Two-step' and 'One-step' so that variables *Outdoor* and *Health* have been used as auxiliary variables. The two-step method is based on the two consecutive models (Models 2b and 3b in Table 3) whereas the one-step method has been constructed direct to the second-phase respondents (Model 4b in Table 3)

Figure 7 presents some comparisons when the two new variables have been added to the response propensity model. The results are quite predictable since this reduces the bias in these estimates and in all other estimates to some extent as well. The bias is still too large in *Boating* and *Biking* in the opinion of many users, I suppose. We can reduce this bias, naturally, by adding new auxiliary variables to the model. How far could we go in this? This has not been examined further in this study. On the other hand, we have worse tools for reducing bias in such variables that have been based on the second survey only. We tested several such estimates and observed some changes in corresponding estimates, being of the same level as in the cases of *Boating* and *Biking* in Figure 7. In this case, however, we cannot check the bias. We can only believe basing on our previous exercises that these results are less biased than those based on more poorly adjusted ones.

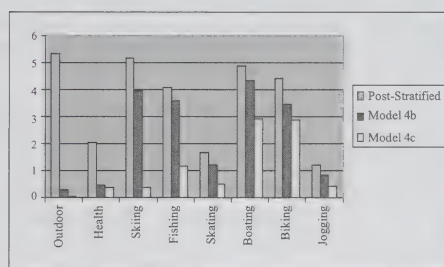


Figure 7 Bias in estimates in percentage points for respondents after both steps based on post-stratification, and the two 'One-step' response propensity methods so that variables *Skiing* and *Fishing* have also been used as auxiliary variables (Model 4c in Table 3). These are compared to those based on Model 4b

5. Discussion

The problem discussed in this paper is common in surveys. There are many surveys which are conducted in more than one step, and some inconsistencies have occurred between these surveys due to missingness and other discrepancies. An internationally well-known example is the European Social Survey (ESS) that includes two questionnaires, a core one and a supplementary one. The number of respondents is naturally smaller for the latter than for the former. This leads to some selectiveness, for example, responding to the second questionnaire being positively associated to political activity. This is awkward from the user's point of view because an estimate based on a larger dataset differs from that based on a smaller one, although both concern the same variable and time period.

Similarly to the ESS, this study concerns two-phase surveyed data. The response rate in the second survey was substantially lower than in the ESS. The effect from selectiveness is also higher. Using the response propensity models we predicted this selectiveness and exploited the results in weighting adjustments, and as the final step we calibrated the sums of the weights to correspond to certain known population aggregates. This strategy aims at making the most of all available auxiliary information, derived both from registers and other external sources, and also of the previous phase of the survey at the micro level.

In our example, the second phase of the survey comprised two different steps but only one data collection. The first step concerned willingness to participate voluntarily in the second phase of the survey, and the second step the actual survey participation of these volunteers, respectively. We examined both steps separately and found interesting information on their response mechanisms. Moreover, we used the results from this analysis for reweighting adjustments. For the sake of comparison, we looked at these both steps in one occasion and built a respective model, and continued the reweighting analogously. Finally, we compared the estimates. It was somewhat surprising that the two results differed quite little in our examples. This is, on the other hand, a good point, since it is easier to work with one step, and hence this could be introduced into use.

We thus propose a certain methodology for two-phase sampling weighting, but cannot say definitely which specification would be the best in each particular case. Our methodology is quite easy to exploit, but the advantages from it depend naturally largely on the availability of good external and internal auxiliary data. If no direct auxiliary variables are available, it will not be clear how good the adjusted estimates will be. Our examples show that these will be easily less biased than the initial ones. However, our recommended technique seems to be somewhat conservative so that all the best adjusted estimates in our analysis are slightly overestimated although not statistically significantly. This is an interesting question for future research that is still needed especially because this problem is becoming more common in the survey world. Another interesting topic for future research is how to make an optimal choice of auxiliary variables in the two-phase survey setting.

Acknowledgements

The author would like to thank the editor and the anonymous referees for their precise and helpful comments.

References

- Déville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Duncan, K.B., and Stasny, E.A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27, 2, 121-130.
- Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-135.
- Ekholm, A., and Laaksonen, S. (1991). Weighting via response modelling in the Finnish household budget survey. *Journal of Official Statistics*, 7, 2, 325-337.
- Fuller, W.A., Loughin, M.M. and Baker, H. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Laaksonen, S. (1999). Weighting and auxiliary variables in sample surveys. In "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches" (Eds., G. Brossier and A.-M. Dussaix). Dunod, Paris. 168-180.
- Laaksonen, S. (2006a). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications, An International Journal*. Publisher: IOS Press. 1, 95-100.
- Laaksonen, S. (2006b). Need for high quality auxiliary data service for improving the quality of editing and imputation. In *United Nations Statistical Commission, "Statistical Data Editing"*, 3, 334-344.
- Laaksonen, S., and Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics*, 81-95.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Lundström, S., and Särndal, C.-E. (2001). Estimation in the presence of nonresponse and frame imperfections. *Statistics Sweden*.
- Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
- Virtanen, V., Pouta, E., Sievänen, T. and Laaksonen, S. (2001). Luonnon virkistyskäytön kysyntätutkimuksen aineistot ja menetelmät. (Data and methods of survey on recreational use of nature). In Luonnon Virkistyskäyttö (Recreational use of nature). Finnish Forest Research Institute, 802.
- Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Weighting in rotating samples: The SILC survey in France

Pascal Ardilly and Pierre Lavallée¹

Abstract

The European Union's Statistics on Income and Living Conditions (SILC) survey was introduced in 2004 as a replacement for the European Panel. It produces annual statistics on income distribution, poverty and social exclusion. First conducted in France in May 2004, it is a longitudinal survey of all individuals over the age of 15 in 16,000 dwellings selected from the master sample and the new-housing sample frame. All respondents are tracked over time, even when they move to a different dwelling. The survey also has to produce cross-sectional estimates of good quality.

To limit the response burden, the sample design recommended by Eurostat is a rotation scheme consisting of four panels that remain in the sample for four years, with one panel replaced each year. France, however, decided to increase the panel duration to nine years. The rotating sample design meets the survey's longitudinal and cross-sectional requirements, but it presents some weighting challenges.

Following a review of the inference context of a longitudinal survey, the paper discusses the longitudinal and cross-sectional weighting, which are designed to produce approximately unbiased estimators.

Key Words: Longitudinal survey; Panel; Weight share method; Longitudinal weighting; Cross-sectional weighting.

1. Introduction

Statistics on Income and Living Conditions (SILC) is a European survey that produces data on the income and living conditions of persons living in regular households (persons living in communal households are excluded). It was introduced in 2004 as a replacement for the European Panel. While it is a European Union (EU) survey and therefore under Eurostat responsibility, it is conducted independently in each EU member state. Hence, the member states - France in this case - are free to adjust the sample design suggested by Eurostat to meet their national requirements. The data are also processed by the individual member states, as is usually the case for Eurostat surveys in the EU. This article deals only with the SILC survey conducted in France, but it may also be of interest to other EU member states.

SILC is a *longitudinal survey* conducted once a year in May. It focuses on individuals rather than households, and data are collected through personal interviews with every person in the sampled dwellings. SILC can be thought of as the European version of the Statistics Canada's Survey of Labour and Income Dynamics (SLID) (see Lavallée 1995, and Lévesque and Franklin 2000).

The SILC sample is rotating: each year, it is formed by combining nine panel subsamples selected under identical steady-state conditions, partly from the master sample and partly from the new-housing sample frame. The master sample and the new-housing sample frame are two dwelling frames constructed from the French census of population

and the information and automated data processing system for dwelling and office space (SITADEL) respectively (see Ardilly 2006).

Each incoming panel includes all individuals living in the selected dwellings. Surveying all members of the households living in the selected dwellings makes it possible to produce both individual-level and household-level estimates and helps keep collection costs down by maximizing the number of individuals reached in each contact. On the other hand, some of the estimates are narrower in scope, applying only to the population aged 16 and over on December 31 of the survey year.

Each year, one subsample is rotated out and replaced with another subsample. In the survey's starting year, 2004, each subsample consisted of 1,780 dwellings (give or take a few units because of rounding). In the second and subsequent years (*i.e.*, from 2005 on), the size of the year's incoming subsample was 3,000 dwellings. Note that at the outset in 2004, the sample was 16,000 dwellings, divided into nine equal parts. One of those parts was surveyed only once (in 2004), another twice (2004 and 2005), a third three times (2004, 2005 and 2006), and so on. After the start-up phase, a given panel will be surveyed for nine consecutive years. During the start-up phase, which will end in 2012 with the departure of the ninth and last subsample from the 2004 selection, the subsamples will have been surveyed fewer than nine times.

The sampling procedure itself is the standard method of selecting units from the master sample and the new-dwelling sample frame (see Ardilly 2006). In this case, no

1. Pascal Ardilly, Sampling and statistical data division, INSEE, Paris, France. E-mail: pascal.ardilly@atih.sante.fr; Pierre Lavallée, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 Canada. E-mail: pierre.lavallee@statcan.ca.

category of individuals is overrepresented. The survey has a uniform sampling fraction - ignoring rounding - except for vacant rural dwellings and dwellings that were secondary residences in the 1999 census and became principal residences by survey date, which are traditionally under-represented.

Under the collection process, each subsample is considered a true panel of individuals. Panel members who move are tracked, and their files are sent to the appropriate regional branch of INSEE. More details on SILC's sample design are available in the November 17, 2003, issue of the *Official Journal of the European Union* and internal INSEE documents describing sampling practices in France.

Since SILC is a longitudinal survey with panels that overlap in time, weighting the sample presents a special problem. This paper provides a detailed picture of the two types of weighting used for SILC. We will begin by discussing some general principles related to SILC's sample design. Then we will examine longitudinal weighting, followed by cross-sectional weighting.

Note that we will not consider the topics of non-response correction and estimate adjustment. Those issues are dealt with in the same way as they are generally for any other longitudinal survey, such as the SLID (see Lavallée 1995, and Lévesque and Franklin 2000).

2. General principles

2.1 Two approaches: The longitudinal view and the cross-sectional view

Each year, we have a sample of fully panelized individuals, eight ninths of whom were interviewed at least once in previous years (barring non-response).

Two types of parameters may be of interest: annual totals Y_t (or their satellites), and changes in totals $\Delta_{t,t+1}$ between two years, consecutive or otherwise. For simplicity, we will confine ourselves to differences in totals between two consecutive years. When discussing changes, we have to be clear about the inference populations involved. We can look at the data in two different ways: either as populations that change over time - the cross-sectional approach - or as a fixed population - the longitudinal approach. If we let Ω_t be the entire in-scope population in year t , the annual total for year t is given by $Y_t = \sum_{i \in \Omega_t} Y_i^t$, where Y_i^t is a variable of

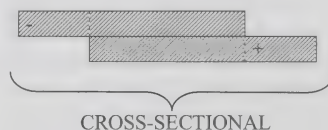
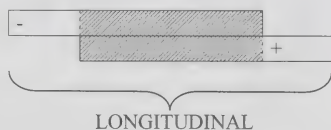
interest measured for individual i . When we look at change, we may want to estimate the difference $\Delta_{t,t+1}^*$ between the total Y_{t+1} at $t+1$ over Ω_{t+1} and the total Y_t at t over Ω_t , that is, $\Delta_{t,t+1}^* = Y_{t+1} - Y_t$. This is a cross-sectional view. Alternatively, we may want to estimate the difference $\Delta_{t,t+1}$ between the totals for the units that are common to populations Ω_{t+1} and Ω_t , where the size difference between the two populations is due to their incoming units (births) and outgoing units (deaths). This is a longitudinal view. Let $\Omega_{t,t+1} = \Omega_t \cap \Omega_{t+1}$, the population that is common to t and $t+1$. Then $\Delta_{t,t+1}$ is defined as $\Delta_{t,t+1} = \sum_{i \in \Omega_{t,t+1}} (Y_i^{t+1} - Y_i^t)$.

The two approaches are illustrated in the diagrams below. The upper rectangle represents the entire population at time t , and the lower one represents the entire population at time $t+1$. The “minus” side represents deaths in the broad sense (persons who have died, emigrated, moved to a communal household, and so on), and the “plus” side represents births in the broad sense (newborns, new immigrants, persons who have become part of the survey population by passing an age threshold, and so on). The grey portion represents the inference population on each date.

2.2 Surveys repeated over time and potential strategies

The goal, of course, is to produce both longitudinal estimates and cross-sectional estimates. There are essentially three possible strategies:

1. An “independent” sampling each year. In fact, because we have a master sample and a new-housing sample frame, the panels are selected from the same localities each year, and as a result, the subsamples are not truly independent. This solution can be improved for estimating changes.
2. A fully panelized sampling, *i.e.*, initial selection of a sample that is surveyed each year. This scenario presents a response burden problem, since the SILC survey is to continue indefinitely. It is therefore unrealistic.
3. A rotational sample. This is the scenario that was chosen, because of its advantages in satisfying both longitudinal and cross-sectional goals.



The table below characterizes the three potential sample designs in terms of the two desired approaches.

Sample TYPE	CROSS-SECTIONAL approach	LONGITUDINAL approach
"Independent" each year	CUSTOMARY	POSSIBLE but less efficient
Panel	IMPOSSIBLE without a top-up sample	CUSTOMARY
Rotational	POSSIBLE	POSSIBLE

The rotation strategy has four major advantages:

- It reduces the sampling error associated with measuring change (in principle, as do panels, though it is theoretically less efficient than a "pure" panel).
- It has a smaller response burden than a "pure" panel. Under the circumstances, since France has a nine-year panel, this argument must be used with restraint. It is more persuasive in the scenario recommended by Eurostat, which consists of an annual survey for four consecutive years.
- It takes into account very "naturally" how the population changes over time. This point will become clearer when we look at the coverage of new populations.
- It reduces observation errors (as do panels).

On the other hand, the strategy also has at least three weaknesses:

- Participants have to be tracked over time, which results in tracing costs and non-response due to moves.
- The length of the individual series is limited to nine years, which is substantial, though not as informative as a pure panel.
- The longitudinal/cross-sectional weighting method is not straightforward.

3. Longitudinal weighting

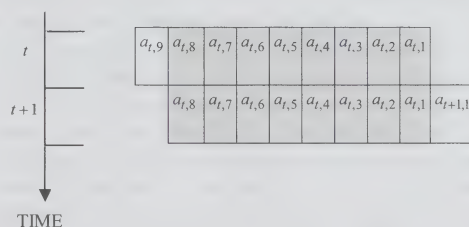
This type of weighting is inherently somewhat easier to understand than cross-sectional weighting because there is no need to take account of how the population changes over time (except for "deaths", units that leave the survey population over time, but they are not much of a technical problem). The idea behind longitudinal estimation, of course, is to make an inference based on a single population at an initial date.

Clearly, the rotational nature of the sample design is what makes weighting difficult, since between two consecutive

years t and $t+1$, we have to deal with eight different panels, each selected from a different population (a population made up of individuals who are, naturally, different from year to year). If we were dealing with just one panel, we would only need to use the sampling weights associated with the panel members who were still in-scope on date t , since those weights are calculated once and for all at the time of selection and can be used to make inferences about the initial population each year throughout the panel's life.

The essential difficulty is to represent population Ω_t on date t using eight panel subsamples selected on different dates and therefore from different populations. Intuitively, it makes sense that a given individual would have a probability of selection on date t that would depend on the number of panel subsamples for which he or she could be chosen. For this discussion, it is assumed that there is no non-response. This situation can be expressed formally by letting $a_{t,k}$ = a panel subsample to be surveyed in year t for the k^{th} time, and $s_{t,t+1} = \bigcup_{k=1}^8 a_{t,k}$.

Note that we can write $a_{t+1,k+1} = a_{t,k}$ ($\forall t, \forall k \neq 9$) since we are obliged to use each (non-outgoing) panel subsample in its entirety year after year. This is pictured below.



The grey part represents $s_{t,t+1}$, which is the sample used in this longitudinal approach. It is from the individuals in $s_{t,t+1}$ that we obtain both Y_i^t and Y_i^{t+1} , i.e., information about individual i on dates t and $t+1$ respectively.

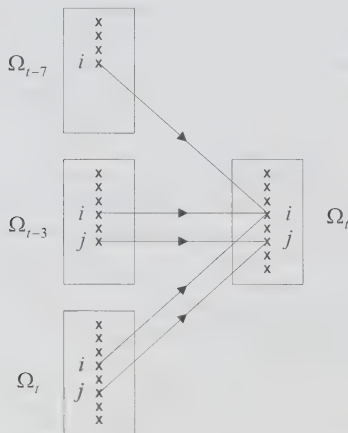
Suppose we have an individual i in Ω_t who is in-scope on date t . We denote as L_i the number of years in $\{t-7, t-6, \dots, t-1, t\}$ during which individual i was in-scope and therefore had a chance of being selected as a member of an incoming panel. It is assumed here that each year, the sample frame covers the survey population exactly. We have $L_i \in \{1, 2, 3, \dots, 8\}$. In addition, we denote as K_i the set of k indexes out of $1, 2, 3, \dots, 8$ for which $i \in a_{t,k}$. These are the numbers of the panels of which individual i is a member on date t . For all i in $s_{t,t+1}$, K_i will be construed as a set containing at least one element. Most of the time, K_i will in fact have only one index, but in some cases, it will have two or even more indexes. That will be the case if i is selected for a panel, he/she moves and his/her new dwelling is chosen for another panel in a subsequent year. Note that our scenario excludes the possibility of selecting a

given dwelling twice, since dwellings from the master sample and the new-housing sample frame are not supposed to be surveyed more than once. This is just a practical convention, however, as the theory can easily accommodate a system in which dwellings can be selected multiple times.

If $i \in a_{t,k}$, let $W_i(t, k)$ be his/her “raw” sampling weight. In fact, it is the sampling weight of the dwelling in which i was living at the time he/she was chosen as a panel member, *i.e.*, at the time of the annual selection from Ω_{t-k+1} . This weighting system allows direct inference from subsample $a_{t,k}$ to the entire population Ω_{t-k+1} . In particular, $\sum_{i \in a_{t,k}} W_i(t, k)$ provides an unbiased estimate of the total number of in-scope individuals who are members of population Ω_{t-k+1} . For SILC in France, that total is roughly 60 million. The longitudinal weight assigned to each individual i in $s_{t,t+1}$ will therefore be as follows:

$$W_i^{t,t+1} = \frac{1}{L_i} \sum_{k \in K_i} W_i(t, k). \quad (1)$$

This equation is derived from the application of the weight share method (see Lavallée 1995, and Lavallée 2002) in which the initial population (the population of sampling units) is defined as the union of the populations $\Omega_{t-7}, \dots, \Omega_{t-1}, \Omega_t$ and the final population (the population of observation units) as Ω_t . This is illustrated in the diagram below; for greater clarity, only three of the initial subpopulations are shown. Clearly, the number of links is equal to L_i (in this case, i has exactly eight links, while j must have fewer than eight because it does not appear in the oldest sample frames). In practice, it is realistic to proceed as if $\Omega_{t-7} \subset \Omega_{t-6} \subset \dots \subset \Omega_{t-1} \subset \Omega_t$. We can work with nested populations, since all individuals who leave the survey population in the time before t will not be part of $s_{t,t+1}$.



Equation (1) is the most general formula for the “raw” longitudinal weight. It can then be simplified for specific situations. For example, if we ignore the cases in which a panel member can be selected more than once, we have

$$W_i^{t,t+1} = \frac{W_i}{L_i} \quad (2)$$

where W_i is the weight of i relative to the one panel subsample of which he/she was a member on date t . In France’s case, because of the sample sizes involved, it seems quite appropriate to use that equation. If we assume that we are in the ideal position - though that seems simplistic in our circumstances - of having a population that does not change over time, we will have $L_i = 8$ for all i . The population changes a great deal in nine years, but with shorter panel lives, the ideal case might be an acceptable approximation. Moreover, if the panels are selected with equal probabilities, W_i will be equal to a constant W and we will have

$$W_i^{t,t+1} = \frac{W}{8}. \quad (3)$$

Such a scenario is highly improbable in France’s case. First, up to 2012, the sample will contain subsamples with very different raw weights. Second, the sampling process is likely to focus on generating a predetermined number of dwellings (as the total number of dwellings increases), and not a constant sampling fraction.

Note that equation (3) is intuitive. Ultimately, everything proceeds “as if” any individual in the longitudinal sample $s_{t,t+1}$ had a selection probability eight times the selection probability of each panel subsample $s_{t,t+1}$ that is part of.

The foregoing applies to the survey in its steady state and must be adapted slightly during the start-up phase, *i.e.*, until 2012. The first longitudinal operation is performed on the combined 2004-2005 data, to estimate the changes between 2004 and 2005 with the 2004 reference population (from which the “deaths” are removed in 2005). In this case, we need only to divide all the weights W_i of the eight subsamples $a_{2004,1}$ to $a_{2004,8}$ by 8; in other words, $L_i = 8$ for all i . In 2006, when we look at the 2005-2006 changes, the denominator L_i may take only two values. In the first scenario, panel member i was in the sample frame used in 2004 (and hence could have been selected in 2004) and so $L_i = 8$. This is due to the fact that everything proceeds as if, in 2004, the seven selection processes for panels $a_{2005,2}$ to $a_{2005,8}$ had been carried out under exactly the same conditions. In the second scenario, individual i was not in the 2004 sample frame - but is in the 2005 frame and is necessarily in $a_{2005,1}$ - and $L_i = 1$. For the 2006-2007 changes, L_i can be equal to 1, 2 or 8, and so on. We will not

have the set of all possible values of L_i in $\{1, 2, 3, \dots, 8\}$ until we measure the 2011-2012 changes.

Once we reach this point in the longitudinal weighting process, we can calculate the longitudinal weights $W_i^{t,t+1}$ and then derive the estimator of the difference $\Delta_{t,t+1}$ using

$$\hat{\Delta}_{t,t+1} = \sum_{s_{i,t}} W_i^{t,t+1} \cdot (Y_i^{t+1} - Y_i^t). \quad (4)$$

Logically, the weights $W_i^{t,t+1}$ are used only to estimate change. They are of no value for point estimates because the inference population has little meaning on a particular date. Note that up to this point, the $W_i^{t,t+1}$ have not been corrected for non-response or adjusted in any other way. In practice, equation (4) will be subject to adjustments in the case of the SILC survey.

Estimation of the difference $\Delta_{t,t+1}^* = Y_{t+1} - Y_t$ is a cross-sectional matter and therefore involves the weighting process described in the next section.

4. Cross-sectional weighting

The aim is to make an inference about the total in-scope population Ω_t on the current date t . The essential difficulty lies in the fact that in theory, a given (panelized) subsample provides adequate coverage of the population only in the year in which it was selected. After that year, the panel subsample no longer represents the new population of “births”, the units that become in-scope. That is the case for newborns, immigrants, individuals who reach specific age thresholds, homeless people who start living in a regular dwelling, people who leave communal dwellings, and so on. While in practice we might consider this coverage defect acceptable for a period of time, it very quickly becomes a serious problem (that is true each year for most panel subsamples), and a top-up sample must be obtained in some fashion. It is worth noting that the problem of population change over time is highly dissymmetrical, since the subpopulation that disappears from year to year (the “deaths”) presents no particular difficulties for weighting.

In the SILC survey, the top-up sample is obtained as follows. We survey all individuals in the household of each panel member interviewed in the longitudinal tracking process. Thus, every household surveyed in the cross-sectional process is made up of two types of people: panel members and cohabitants (people who are surveyed but are not panel members). This method covers a large portion of the “births” (in the broad sense) in the population. However, it does not cover households consisting entirely of “births”, such as households of new immigrants. “Birth” status is usually determined by asking the birth date of newborns and the landing date of immigrants. Moreover, in practice, the

weakness in births coverage is generally regarded as very minor because it is partially corrected with adjustments.

The main technique used to produce cross-sectional weights is the weight share method (Lavallée 2002). As noted previously, in year t we have nine panel subsamples $a_{t,k}$ ($1 \leq k \leq 9$). We will describe two different ways of using the weight share method. The information that must be collected in the questionnaire is the same for both methods.

4.1 Method 1

The more rigorous approach involves linking all nine subsamples $a_{t,k}$ to the cross-sectional sample for year t , which we will denote \tilde{u}_t (Merkouris 2001). In other words, the sample \tilde{u}_t is the same as $s_{t,t} = \bigcup_{k=1}^9 a_{t,k}$. First, we must determine the links associated with this approach. When a panel member in one of the nine subsamples $a_{t,k}$ is selected, he/she points to himself/herself as a member of the cross-sectional sample at t (similar to what is shown in the diagram in 3.1). Under these conditions, when the survey is in steady-state mode, the cross-sectional weight $W_i^{t(1)}$ of an individual i in \tilde{u}_t is calculated as shown below. The household of which i is a member is denoted m . We have

$$W_i^{t(1)} = \frac{\sum_{k=1}^9 \sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t, k)}{\sum_{k=1}^9 \sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \quad (5)$$

where $W_j(t, k)$ is the sampling weight from sample $a_{t,k}$.

This expression shows that all members of the same household ultimately have the same weight. In the numerator, we have the sum of all the “raw” weights (the sampling weights) of the household’s panel members. It is understood that a panel member generally appears in only one subsample, but that there may be cases in which a panel member is selected two or more times over a period of nine consecutive years (usually because he/she has moved). Note that dwellings selected from the master sample and the new-housing sample frame are not supposed to be selected again and therefore, in the case of SILC, the probability that an individual who has not moved will appear in two different panels is zero.

As in the longitudinal case (see 3.1), weighting can be carried out only if the data management system is capable of linking each panel member in \tilde{u}_t to all panel samples $a_{t,k}$ in which he/she is included. In the denominator, for each of the nine years $t - 8$ to t considered, we count the household members (both panel members and cohabitants) who are in the sample frame from which the incoming panel subsample

for the year in question is selected. This calculation clearly requires the information provided by the questionnaire.

There are two advantages to this approach: it is completely general, and it produces unbiased cross-sectional weights directly because every cross-sectional household is necessarily linked to one of the nine subsamples involved. The fact that there is an incoming subsample each year ensures the completeness of the cross-sectional population Ω_t ; that is, in more technical terms, it ensures that there is at least one link for each household considered at t . This is a useful property of rotational sampling, as discussed in section 2.2. On the other hand, the weighting formula has a disadvantage, which is its (relative) complexity both in theoretical terms and for computer programming purposes.

In the start-up phase (up to and including 2011), the formula must be adjusted. The numerator remains the same, but the denominator covers all individuals who could be sampled in 2004 (the survey's first year) and subsequent years. In 2004, weighting is trivial since there is no weight share, but in 2005, we have

$$W_i^{t(1)} = \frac{\sum_{k=1}^9 \sum_{j \in m} W_j(t, k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 8 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \quad (6)$$

In 2006, the formula will be

$$W_i^{t(1)} = \frac{\sum_{k=1}^9 \sum_{j \in m} W_j(t, k)}{\left(\sum_{\substack{j \in m \\ j \in \Omega_{2006}}} 1 \right) + \left(\sum_{\substack{j \in m \\ j \in \Omega_{2005}}} 1 \right) + 7 \cdot \left(\sum_{\substack{j \in m \\ j \in \Omega_{2004}}} 1 \right)}. \quad (7)$$

4.2 Method 2

It is possible to take an alternative approach to cross-sectional weighting one that leads to a "slightly" simpler equation and is easier to program, but one that presents a difficulty that was not present in the previous method and may make the final weights somewhat less precise. The idea is to use one subsample at a time rather than all of them at once. We take one of the nine subsamples $a_{t,k}$ and the sample of households to which it leads. We then apply the weight share, which when the survey is in steady-state mode yields an individual weight equal to

$$\tilde{W}_j(t, k) = \frac{\sum_{\substack{j \in m \\ j \in a_{t,k}}} W_j(t, k)}{\sum_{\substack{j \in m \\ j \in \Omega_{t-k+1}}} 1} \quad (8)$$

for any individual i in household m . It is very easy to verify that if $k = 1$ (which is the case for the incoming subsample), $\tilde{W}_j(t, 1)$ is the sampling weight of household m .

The problem with this approach lies in the existence (*a priori*) on date t of individuals who cannot be surveyed because they belong to households that cannot be "reached" through the sampling $a_{t,k}$ (as long as $k \geq 2$), i.e., individuals whose probability of being surveyed at t is zero. This impediment did not exist in the previous method because taking all the subsamples into account at once ensured that on date t , every household had a non-zero probability of being selected, at least through $a_{t,1}$. This illustrates once again one of the key advantages of rotational sampling, which is that it covers the entire population each year. In our approach, it is clear that if we consider $a_{t,k}$ for $k \geq 2$, the population of households consisting exclusively of "immigrants" (in the broad sense) between $t - k + 1$ and t is not covered. To formalize the situation and produce the final cross-sectional weight, we will use $\Omega_{\alpha,t}^{\text{immig}}$ to denote the population of "immigrants" (in the broad sense) on date t in households that consist only of immigrants who can be sampled after year α , with $t - 8 \leq \alpha \leq t - 1$. To be more precise, we should say "who can be sampled on or after a date that is strictly subsequent to the collection date in year α ."

On date t , the entire population Ω_t is partitioned into nine components: the eight subpopulations $\Omega_{\alpha,t}^{\text{immig}}$, with α ranging from $t - 8$ to $t - 1$, and the subpopulation consisting of individuals who either were already surveyable at $t - 8$ or became surveyable on a date subsequent to $t - 8$ (i.e., who immigrated after $t - 8$) but at t are members of a household containing at least one person who is surveyable at $t - 8$. We consider that if the household at t contains at least one person who is surveyable at $t - 8$, that will be the case on any date between $t - 8$ and $t - 1$. This ignores situations in which an individual who is in-scope on a given date becomes out-of-scope for a time (as a result of emigration, for example) and then becomes in-scope again.

Next, we use $\tilde{u}_{t,k}$ to denote the cross-sectional sample at t from panel $a_{t,k}$, which leads to $\bigcup_{k=1}^9 \tilde{u}_{t,k} = \tilde{u}_t$. Let $Y_{\alpha,t}^{\text{immig}}$ be the total of the Y_i^t defined on $\Omega_{\alpha,t}^{\text{immig}}$. Following the weight share performed for all $k = 2, \dots, 9$, we have

$$\begin{aligned} E \left(\sum_{j \in \tilde{u}_{t,k}} \tilde{W}_j(t, k) \cdot Y_j^t \right) &= \sum_{\Omega_t} Y_j^t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}} \\ &= Y_t - \sum_{\alpha=t-k+1}^{t-1} Y_{\alpha,t}^{\text{immig}} \end{aligned} \quad (9)$$

and

$$E\left(\sum_{j \in \tilde{u}_i} \tilde{W}_j(t, 1) \cdot Y_j^t\right) = \sum_{\Omega_t} Y_j^t = Y_t \quad (10)$$

since $\tilde{u}_{t,1} = a_{t,1}$.

If we were using shorter-duration panels, we might be able to ignore the $Y_{\alpha,t}^{\text{immig}}$ and take the actual total over Ω_t . In that case, the “raw” final cross-sectional weight of any individual i would be $\tilde{W}_i(t, k)/9$ if i is from $a_{t,k}$, which would yield the final estimator

$$\begin{aligned} \hat{Y}_t &= \frac{1}{9} \sum_{k=1}^9 \sum_{i \in \tilde{u}_{t,k}} \tilde{W}_i(t, k) \cdot Y_i^t \\ &= \frac{1}{9} \sum_{i \in \tilde{u}_t} \tilde{W}_i(t, k) \cdot Y_i^t. \end{aligned} \quad (11)$$

However, since the panels used in France have long lives, we will probably not be able to ignore the $Y_{\alpha,t}^{\text{immig}}$ (an analysis of the collection files will provide the answer), which will mean having to compute specific weights for the individuals in $\Omega_{\alpha,t}^{\text{immig}}$. In those circumstances, we check that any individual i in $\Omega_{\alpha,t}^{\text{immig}}$ who ends up in the cross-sectional sample \tilde{u}_t will have a raw cross-sectional weight $\tilde{W}_i^{t(2)}$ equal to the weight share value $\tilde{W}_i(t, k)$ divided by $t - \alpha$ (and therefore $1 \leq t - \alpha \leq 8$). Any individual in Ω_t who does not belong to any of the $\Omega_{\alpha,t}^{\text{immig}}$ (i.e., the vast majority of individuals) will have a final weight of $\tilde{W}_i(t, k)/9$. Note that if i is in $\Omega_{\alpha,t}^{\text{immig}}$, he/she can be surveyed only through $a_{t,1}, a_{t,2}, \dots, a_{t,t-\alpha}$. Thus we have

$$W_i^{t(2)} = \begin{cases} \tilde{W}_i(t, k)/(t - \alpha) & \text{if } i \in \Omega_{\alpha,t}^{\text{immig}} \\ \tilde{W}_i(t, k)/9 & \text{otherwise} \end{cases} \quad (12)$$

In the start-up phase, the weighting process has to be adjusted. In 2005, the final cross-sectional weight of individuals in $\Omega_{2004,2005}^{\text{immig}}$ will come directly from the selection of the dwelling from $a_{2005,1}$ (they can only be reached through this incoming panel). In contrast, all other individuals can be surveyed “normally” in the nine panels $a_{2005,k}$ ($1 \leq k \leq 9$), so that their weights as calculated by the weight share method will all be divided by 9. In 2006, the weights of the individuals in $\Omega_{2005,2006}^{\text{immig}}$ will be equal to the weight of the dwelling in which they live, a weight that directly reflects the sampling from $a_{2006,1}$; the weights of the individuals in $\Omega_{2004,2006}^{\text{immig}}$ will be the weights from the

weight share divided by 2; and the weights of all other individuals will be the weights from the weight share divided by 9.

This procedure can be carried out for one subsample after another and does not have to take account of what happens in other subsamples. If an individual is surveyed at t through two (or more) different subsamples $a_{t,k}$, we carry out the full procedure for each of the two (or more) subsamples. This could occur, for example, in the case of a household composed of two panel members from two different subsamples $a_{t,k}$ who married each other and before their marriage were each tracked separately as one-person households. In that scenario, each individual would be “formally” surveyed twice, once as a panel member and once as a cohabitant.

Finally, to estimate the difference $\Delta_{t,t+1}^* = Y_{t+1} - Y_t$, we can use the weights $W_i^{t(1)}$ from method 1 and calculate

$$\hat{\Delta}_{t,t+1}^* = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(1)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(1)} Y_i^t. \quad (13)$$

Alternatively, we can use the weights $W_i^{t(2)}$ from method 2. In that case, the estimator of the difference $\Delta_{t,t+1}^*$ will be given by

$$\hat{\Delta}_{t,t+1}^* = \sum_{i \in \tilde{u}_{t+1}} W_i^{t+1(2)} Y_i^{t+1} - \sum_{i \in \tilde{u}_t} W_i^{t(2)} Y_i^t. \quad (14)$$

References

- Ardilly, P. (2006). *Les techniques de sondage*, 2nd edition. Editions Technip, Paris.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Editions de l'Université de Bruxelles et Editions Ellipses.
- Lévesque, I., and Franklin, S. (2000). Longitudinal and Cross-Sectional Weighting of the Survey of Labour and Income Dynamics, 1997 Reference year. Research document on income, Statistics Canada, Catalogue No. 75F0002MIE-00004, June 2000.
- Merkouris, T. (2001). Cross-sectional estimation in multiple-panel household surveys. *Survey Methodology*, 27, 171-181.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). *Méthodologie de l'Enquête sur la population active*. Statistics Canada, Catalogue 71-526.

Cell collapsing in poststratification

Jay J. Kim, Jianzhu Li and Richard Valliant¹

Abstract

Poststratification is a common method of estimation in household surveys. Cells are formed based on characteristics that are known for all sample respondents and for which external control counts are available from a census or another source. The inverses of the poststratification adjustments are usually referred to as coverage ratios. Coverage of some demographic groups may be substantially below 100 percent, and poststratifying serves to correct for biases due to poor coverage. A standard procedure in poststratification is to collapse or combine cells when the sample sizes fall below some minimum or the weight adjustments are above some maximum. Collapsing can either increase or decrease the variance of an estimate but may simultaneously increase its bias. We study the effects on bias and variance of this type of dynamic cell collapsing theoretically and through simulation using a population based on the 2003 National Health Interview Survey. Two alternative estimators are also proposed that restrict the size of weight adjustments when cells are collapsed.

Key Words: Bias; Combining cells; Coverage error; Poststratification; Under-coverage; Weight trimming.

1. Introduction

Poststratification is a common technique used in survey weighting that can serve to (1) reduce variances or (2) adjust for deficient coverage by the sample of some groups in the target population. In household surveys in the U.S. the second purpose is especially important because some demographic groups, like young Black males, are covered less well than others (*e.g.*, see Kostanich and Dippo 2000, chapter 16). Adjusting for undercoverage can lead to differential weights, which may correct for bias but will also increase standard errors. Practitioners often avoid making extreme weight adjustments, in effect trading-off some bias reduction in order to keep variances under control.

One method of controlling the size of weight adjustments is to collapse the initial poststratification cells together if the adjustment in a cell exceeds some limit. Little (1993) and Lazzeroni and Little (1998) cover methods of collapsing categories of ordinal poststratifiers. Other strategies for how to collapse strata or construct estimators have been suggested by Fuller (1966), Kalton and Maligalig (1991), and Tremblay (1986). Kim, Thompson, Woltman, and Vajs (1982) give some practical applications. In this paper, we study the effects on bias and variance of combining cells, assuming that more finely defined cells would be preferable if the sample sizes and sizes of weight adjustments were within some tolerances set by the survey designers.

Two criteria are often used to decide whether a cell should be collapsed with another. The first is the *inverse coverage ratio* or *initial adjustment factor* (IAF), and is defined as the ratio of the control count to the initially weighted sample count for the cell. A ratio which is significantly different from 1 indicates that coverage is

either low or high for the group represented by the cell. When the IAF for a cell falls outside some bounds set in advance, the cell is combined with another. For example, the collapsing threshold for “high” ratio might be 2 and the threshold for “low” ratio 0.6, which are the bounds used in the Current Population Survey (CPS) conducted by U.S. Bureau of the Census (see Kostanich and Dippo 2000, page 10-7). The second criterion is the sample size. A cell whose raw sample count is too small may be collapsed on the grounds that the IAF is unstable. We will refer to a cell as *sparse* if it violates one or the other of the criteria and is collapsed with another cell.

The categories of the variables that define poststrata are usually sorted based on a natural ordering (*e.g.*, age or income categories) or a convenient ordering (*e.g.*, race-ethnicity). Common practice is to collapse a cell with an adjacent one which is similar in characteristics, disregarding different coverage ratios of the individual cells.

Kalton and Flores-Cervantes (2003, page 95) observed that “methods that automatically restrict the range of the adjustments are redistributing the excess adjustments that would otherwise be given to some respondents to other respondents. The appropriateness of this redistribution should be examined.” This paper indeed examines its appropriateness and identifies circumstances where the weight redistribution due to collapsing may be quite harmful.

An obvious weakness of popular collapsing strategies is that coverage bias for some groups will be incompletely corrected. For example, suppose that the survey estimate for the number of units in a group is only 1/3 of the census count, so that initial weights would have to be multiplied by 3 to correct for undercoverage. If cell collapsing restricts the

1. Jay J. Kim, National Center for Health Statistics, Centers for Disease Control and Prevention; Jianzhu Li, Joint Program in Survey Methodology, University of Maryland; Richard Valliant, Survey Research Center, University of Michigan.

weight adjustment for units in that group to a factor of 2, then the survey estimate for the number of units in the group will be only 2/3 of the census count. In addition, if cells with much different means are combined, bias can be introduced rather than corrected. The incomplete correction for undercoverage and collapsing of cells with disparate characteristics may lead to bias in totals, means, and other types of estimates.

Table 1 gives some illustrative coverage ratios, *i.e.*, survey estimates prior to poststratification divided by census counts, for the March 2002 U.S. Current Population Survey and the 2003 Behavioral Risk Factors Surveillance Survey (BRFSS) in a set of 44 counties in the southwestern U.S. The survey estimates include a nonresponse adjustment for both CPS and BRFSS. Coverage ratios shown for the subset of demographic groups in the CPS range from 0.70 to 0.93 with Black-only typically being less than for other groups. BRFSS is a telephone survey with low response rates in this set of counties, and the ratios for BRFSS are much smaller than for CPS. There are also substantial differences in coverage ratios for different groups in BRFSS. For example, the ratio for 35 - 44 year old Hispanic males is 0.18 but is 0.37 for Black/Multiracial/Other males in the same age range. If these two groups were collapsed, incomplete coverage would be under-corrected for the Hispanics but over-corrected for the Black/Multiracial/Other group. Another example is the 2003 National Health Interview Survey (NHIS) where American Indians and Asians were collapsed with Whites within age groups. In the cell for ages 25-29, for example, the coverage rates for Whites, American Indians, and Asians were 0.60, 0.44, and 0.31, respectively (Tompkins and Kim 2006).

This paper demonstrates the weaknesses of current cell collapsing procedures and proposes some alternatives. Section 2 discusses the bias of some standard estimators when there is undercoverage. Section 3 introduces two new estimators that retain more of the undercoverage adjustment than the standard method when cells are collapsed. Empirical properties of the standard and alternative methods are investigated through simulation in section 4. We conclude in section 5 with a summary and some possibilities for future research.

2. Some standard estimators

Three standard estimators of the population mean are the Hájek estimator, the poststratified estimator, and the poststratified estimator with initial poststrata collapsed together where necessary. Each of these is defined in detail below. When the sampling frame covers units in the target population at different rates, each estimator can be biased. Kim *et al.* (2005) give some numerical illustrations of the

effects of collapsing pairs of cells with different coverage rates.

To derive theory for alternative estimators of means, we model a unit's being covered by the sampling frame and a cell's being sparse or not as random events. Define three indicator variables: $\delta_k = 1$ if unit k is selected for the sample and 0 if not; $c_k = 1$ if unit k is covered by the frame and 0 if not; $d_i = 1$ if poststratum i is classified as sparse in a particular sample and 0 if not ($i = 1, \dots, I$). These indicators are assumed to be mutually independent and to have expectations π_k , ϕ_k , and p_i , respectively. Consider a stratified, two-stage probability sample design. A design stratum is denoted by h ; s_h is the set of primary sampling units (PSU's) selected from design stratum h ; $s_{h(i)}$ is the set of sample units from sample PSU j in stratum h that are also in poststratum i ; U_h is the population of PSU's in stratum h ; U_{hj} is the population of units in PSU j within stratum h ; and $U_{h(i)}$ is the population of units in PSU j within stratum h that are in poststratum i . For the analysis in this section, the sample design does not need to be specified in more detail. Note that some summations in sections 2 and 3 of the form $\sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} \dots$ for units within poststratum i could be simplified to $\sum_{k \in s_{(i)}}$ without loss of generality. We have used the more elaborate notation to make clear how the stages of sampling should be treated.

2.1 Hájek estimator

First, consider the Hájek estimator of a mean, which is

$$\hat{y}_\pi = \frac{\left(\sum_{i=1}^I \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} y_k / \pi_k \right)}{\left(\sum_{i=1}^I \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} 1 / \pi_k \right)} \equiv \hat{T}_\pi / \hat{N}_\pi. \quad (1)$$

The expectation of \hat{T}_π with respect to sampling and the coverage mechanism is $E_c E_\pi(\hat{T}_\pi) = \sum_{i=1}^I \sum_h \sum_{j \in U_h} \sum_{k \in U_{h(i)}} \phi_k y_k \equiv T^c$, where the c superscript denotes "covered". Similarly, the expectation of \hat{N}_π is $E_c E_\pi(\hat{N}_\pi) = \sum_{i=1}^I \sum_h \sum_{j \in U_h} \sum_{k \in U_{h(i)}} \phi_k \equiv N^c$. Expanding \hat{y}_π around (T^c, N^c) , its linear approximation is $\hat{y}_\pi \doteq T^c / N^c + 1/N^c \times (\hat{T}_\pi - (T^c/N^c)\hat{N}_\pi)$. Next, consider the bias of \hat{y}_π as an estimator of $\bar{Y} = \sum_{i=1}^I T_i / N$ with T_i being the total for the full population of units in poststratum i (not just the covered portion). After some calculation, the bias is

$$\text{bias}(\hat{y}_\pi) \doteq \frac{T^c}{N^c} - \frac{1}{N} \sum_{i=1}^I T_i = \frac{C_{\phi y}}{\bar{\phi}} \quad (2)$$

where $\bar{\phi} = \sum \phi_k / N$, $C_{\phi y} = \sum (\phi_k - \bar{\phi})(y_k - \bar{Y}) / N$, $\bar{Y} = T/N$, and \sum denotes the sum over $i, h, j \in U_h$, and $k \in U_{h(i)}$. Consequently, \hat{y}_π is biased if there is any correlation between the variable measured, y , and the coverage probability ϕ_k . The bias in (2) is $O(1)$, meaning that it remains important even in large samples.

Table 1

Coverage ratios from the Current Population Survey (CPS) and Behavioral Risk Factors Surveillance Survey (BRFSS). White only and black only mean that only those races were reported by a respondent in the CPS. Residual-only race group includes cases indicating a single race other than white or black, and cases indicating two or more races. Hispanics may be of any race in the BRFSS tabulation

March 2002 Current Population Survey						
Age	White-only		Black-only		Residual-only	
	Male	Female	Male	Female	Male	Female
0 - 15	0.93	0.93	0.78	0.79	0.91	0.90
16 - 19	0.90	0.88	0.76	0.81	0.93	0.75
20 - 24	0.79	0.85	0.72	0.77	0.75	0.72
25 - 34	0.83	0.89	0.70	0.76	0.76	0.80
All ages	0.90	0.92	0.78	0.83	0.85	0.84
2003 BRFSS: 44 border counties in Arizona, California, New Mexico, and Texas						
Age	White Non-Hispanic		Black, Multiracial, Other		Hispanic	
	Male	Female	Male	Female	Male	Female
18 - 24	0.19	0.26	0.12	0.24	0.15	0.22
25 - 34	0.20	0.31	0.10	0.16	0.19	0.39
35 - 44	0.28	0.31	0.37	0.25	0.18	0.30
All ages	0.25	0.31	0.25	0.20	0.18	0.31

Sources: Bureau of the Census (2002), Gonzalez, Town, and Kim (2005).

If the coverage probability is the same for every unit in poststratum i , i.e., $\phi_k = \phi(i)$ for any $k \in U_{h(i)}$, then the approximate bias reduces to $\text{bias}(\hat{y}_{\pi}) \doteq \phi^{-1} \sum_i W_i \times (\phi(i) - \bar{\phi})(\bar{Y}_i - \bar{Y})$ where $W_i = N_i/N$ and $\bar{Y}_i = \sum_{h,j \in U_h, k \in U_{h(i)}} y_k / N_i$. If there is a correlation between the poststratum coverage probabilities and the poststratum means, the Hájek estimator will again be biased, and the bias could be either positive or negative. If the coverage rates or the means are constant across poststrata, i.e., $\phi(i) = \phi_0$ or $\bar{Y}_i = \bar{Y}$, then the Hájek estimator will be unbiased, but poststrata are usually not formed this way. Also, the bias exists even when the appropriate set of poststrata, that subdivide the population into groups with different means, is unknown to the sampler.

2.2 Poststratified mean with no cell collapsing

The poststratified mean is defined as $\hat{y}_{\text{PSI}} = 1/N \sum_{i=1}^I (N_i / \hat{N}_{\pi i}) \hat{T}_{\pi i}$ where $\hat{T}_{\pi i}$ and $\hat{N}_{\pi i}$ are defined as in (1) but excluding the summation over i . Define $T_i^c = \sum_{h,j \in U_h, k \in U_{h(i)}} \phi_k y_k$ and $N_i^c = \sum_{h,j \in U_h, k \in U_{h(i)}} \phi_k$. These are the expected (with respect to the coverage mechanism) total and count of covered units in poststratum i . Expanding \hat{y}_{PSI} around (T_i^c, N_i^c) , $i=1, \dots, I$, its linear approximation is

$$\hat{y}_{\text{PSI}} \doteq \frac{1}{N} \left[\sum_i \frac{N_i}{N_i^c} T_i^c + \sum_i \frac{N_i}{N_i^c} \left(\hat{T}_{\pi i} - \frac{T_i^c}{N_i^c} \hat{N}_{\pi i} \right) \right].$$

The bias of the poststratified estimator is then the first term of this expression minus $\sum_{i=1}^I T_i^c / N$, and after some manipulation, can be written as

$$\text{bias}(\hat{y}_{\text{PSI}}) \doteq \sum_{i=1}^I W_i \frac{C_{\phi i}}{\bar{\phi}_i} \quad (3)$$

where $\bar{\phi}_i = \sum \phi_k / N_i$, $C_{\phi i} = \sum (\phi_k - \bar{\phi}_i)(y_k - \bar{Y}_i) / N_i$, and \sum denotes the sum over $h, j \in U_h$, and $k \in U_{h(i)}$. Thus, \hat{y}_{PSI} is biased if there is any correlation between the y variable measured and the coverage probability ϕ_k in any of the poststrata. If the coverage rate is constant at $\phi_k = \phi(i)$ within poststratum i , then the poststratified estimator is approximately unbiased. From (3) it is apparent that poststrata should be formed so that either coverage rates or the y 's are homogeneous within each poststratum. This is similar to the recommendations of Eltinge and Yansaneh (1997), Kalton and Maligalig (1991), and Little and Vartivarian (2005) for the formation of nonresponse adjustment cells. In large surveys, the initial set of candidate poststrata is often more extensive than the sample can support. With few exceptions, some of the initial poststrata are collapsed to control weight adjustments. If no collapsing occurs, this is usually because small categories are pre-collapsed based on prior experience in the same or a similar survey. In that sense, PSI does not really exist in practice. More common is the collapsing approach, PS2, described below.

2.3 Poststratified mean with collapsing

Turning to the poststratified estimator with collapsing, the sparse cells are identified and combined with other cells considered to be their nearest neighbors. This could result in more than one sparse cell being collapsed with a given nonsparse cell. Neighbors can be defined in various ways, e.g., cells with similar estimated coverage rates, $\hat{N}_{\pi i} / N_i$, cells that are adjacent in some substantive sense like nearby income classes, or cells that have similar means on some important survey variables. The general algorithm for collapsing, given an initial set of cells, is:

- (1) Compute the collapsing criteria for each cell, e.g., the IAF's, $N_i/\hat{N}_{\pi i}$, and the cell sample sizes;
- (2) Identify the sparse cells, i.e., those whose criteria fall outside the bounds for collapsing;
- (3) Determine the nearest, nonsparse neighbor of each sparse cell and combine the sparse cell with its neighbor.

The poststratified mean with collapsing is then $\hat{y}_{PS2} = 1/N \sum_g (N_g/\hat{N}_{\pi g}) \hat{T}_{\pi g}$ where $\hat{T}_{\pi g} = \sum_{i \in A_g} \sum_{h, j \in s_h, k \in s_{h(i)}} y_k/\pi_k$ and $\hat{N}_{\pi g}$ is defined similarly. Define $T_g^c = \sum_{A_g} T_i^c$ and $N_g^c = \sum_{A_g} N_i^c$ with A_g being the set of poststrata in collapsed group g . Expanding \hat{y}_{PS2} around (T_g^c, N_g^c) , for each collapsed group g , gives

$$\hat{y}_{PS2} \doteq \frac{1}{N} \left[\sum_g \frac{N_g}{N_g^c} T_g^c + \sum_g \frac{N_g}{N_g^c} \left(\hat{T}_{\pi g} - \frac{T_g^c}{N_g^c} \hat{N}_{\pi g} \right) \right].$$

It follows that

$$\text{bias}(\hat{y}_{PS2}) \doteq \sum_g W_g \frac{C_{\phi g}}{\bar{\phi}_g} \quad (4)$$

with

$$W_g = N_g/N, \bar{\phi}_g = \sum \phi_k/N_g, C_{\phi g} = \sum (\phi_k - \bar{\phi}_g)(y_k - \bar{Y}_g)/N_g,$$

and the summations in $\bar{\phi}_g$ and $C_{\phi g}$ are over $i \in A_g, h, j \in U_h$, and $k \in U_{h(i)}$. If ϕ_k is constant within collapsed group g , this estimator is unbiased, but if $\phi_k = \phi(i)$, i.e., the coverage rate is constant within poststratum i but can differ across the poststrata, then the bias becomes

$$\text{bias}(\hat{y}_{PS2}) \doteq \sum_g W_g \frac{C_{\phi g}^*}{\bar{\phi}_g} \quad (5)$$

with $\bar{\phi}_g = \sum W_{gi} \phi(i)$, $C_{\phi g}^* = \sum W_{gi} (\phi(i) - \bar{\phi}_g)(\bar{Y}_i - \bar{Y}_g)$, $W_{gi} = N_i/N_g$, and the summations are over $i \in A_g$.

Thus, in the case where \hat{y}_{PS1} will be unbiased, \hat{y}_{PS2} will be biased if poststrata are collapsed together that have different coverage rates and different population means. Since $\bar{\phi}_g$ and $C_{\phi g}^*$ are both $O(1)$, the bias does not decrease as the sample increases; thus, the bias-squared will eventually be the dominant part of the mean square error. If cells are collapsed, the cells in each group should have the same coverage rates, the same means, or both to avoid bias.

3. Weight restricted estimators

We examine two alternative methods of weight computation when collapsing of poststrata is used, extending work of Kim (2004). The alternatives are designed to be compromises between (a) use of all poststrata and the potential for large weight adjustments and (b) collapsing of poststrata yielding less variable weights but potentially

biased estimates. We refer to these as *weight restriction* (WR) methods. The two alternatives presented in this section use cell collapsing but retain a larger share of the weight adjustment for individual cells than does the standard collapsing method.

The first alternative is denoted PS.WR1 and consists of the following algorithm. Denote the maximum allowable weight adjustment by f_{\max} with $f_{\max} > 1$.

- (1) Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.
- (2) Censor any IAF greater than f_{\max} to f_{\max} and adjust each weight in the corresponding initial cell to $\tilde{w}_k = w_k f_{\max}$ with $w_k = 1/\pi_k$. For units in cells with $\text{IAF} \leq f_{\max}$, set $\tilde{w}_k = w_k$.
- (3) Compute a collapsing adjustment factor (CAF) for a collapsed group g as

$$\tilde{f}_g = N_g / \sum_{i \in A_g} \sum_{h, j \in s_h, k \in s_{h(i)}} \tilde{w}_k.$$

- (4) The final adjusted weight is then $\tilde{w}_k \tilde{f}_g$ for unit k in group g .

This method will reduce the largest values of the final weight adjustment below the without-collapsing adjustments, $N_i/\hat{N}_{\pi i}$, though there may be one or more groups that have CAF's greater than the f_{\max} cutoff. The control total for group g , N_g , is met in the sense that $\sum_{i \in A_g} \sum_{h, j \in s_h, k \in s_{h(i)}} \tilde{w}_k \tilde{f}_g = N_g$ but the control totals for the individual cells in A_g are not.

To analyze the properties of PS.WR1, define $A_{g,sp}$ and $A_{g,ns}$ to be the sets of sparse and nonsparse poststrata in collapsed group g . PS.WR1 can be expressed as

$$\hat{y}_{PSWR1} = \frac{1}{N} \sum_g \frac{N_g}{\hat{N}_{gWR1}} \hat{T}_{gWR1}$$

where

$$\hat{T}_{gWR1} = \sum_{i \in A_{g,sp}} \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} f_{\max} w_k y_k + \sum_{i \in A_{g,ns}} \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} w_k y_k$$

and \hat{N}_{gWR1} has a similar definition with y_k set to 1. The expectation of \hat{T}_{gWR1} over the coverage, sparseness, and sampling mechanisms is $E_c E_{sp} E_{\pi}(\hat{T}_{gWR1}) = T_g^c + (f_{\max} - 1) \times \tilde{T}_g^c$ where $\tilde{T}_g^c = \sum_{i \in A_g} p_i T_i^c$. Likewise, $E_c E_{sp} E_{\pi}(\hat{N}_{gWR1}) = N_g^c + (f_{\max} - 1) \tilde{N}_g^c$ with $\tilde{N}_g^c = \sum_{i \in A_g} p_i N_i^c$. \hat{y}_{PSWR1} can be expanded around the expectations of $(\hat{T}_{gWR1}, \hat{N}_{gWR1})$. After some manipulation, the approximate bias of \hat{y}_{PSWR1} becomes

$$\text{bias}(\hat{y}_{PSWR1}) \doteq \sum_g W_g \frac{C_{\alpha\phi, y, g}}{(\alpha\phi)_g} \quad (6)$$

where

$$(\overline{\alpha\phi})_g = \sum \alpha_i \phi_k / N_g, \alpha_i = 1 + (f_{\max} - 1)p_i,$$

$$C_{\alpha\phi, y, g} = \sum (\alpha_i \phi_k - (\overline{\alpha\phi})_g)(y_k - \bar{Y}_g) / N_g,$$

and the summations are over $i \in A_g, h, j \in U_h$, and $k \in U_{hj(i)}$. In the case of a common coverage probability in poststratum i , i.e., $\phi_k = \phi(i)$, we have

$$(\overline{\alpha\phi})_g = \bar{\phi}_g + (f_{\max} - 1)(\overline{p\phi})_g$$

and

$$\alpha_i \phi(i) - (\overline{\alpha\phi})_g = (\phi(i) - \bar{\phi}_g) + (f_{\max} - 1)(p_i \phi(i) - (\overline{p\phi})_g)$$

where $(\overline{p\phi})_g = \sum_{A_g} W_{gi} p_i \phi(i)$. From this it follows that

$$C_{\alpha\phi, y, g} = C_{\phi y, g}^* + (f_{\max} - 1)C_{p\phi, y, g}$$

with

$$C_{p\phi, y, g} = \sum_{A_g} W_{gi} (p_i \phi(i) - (\overline{p\phi})_g)(\bar{Y}_i - \bar{Y}_g).$$

If the cell means \bar{Y}_i are all equal within a collapsed group, then $C_{\phi y, g}^* = C_{p\phi, y, g} = 0$ and $\hat{y}_{\text{PS.WR1}}$ will be approximately unbiased. In the special case in which coverage is constant in a group, i.e., $\phi(i) = \bar{\phi}_g$, then $C_{\alpha\phi, y, g} = (f_{\max} - 1) \bar{\phi}_g \sum_{i \in A_g} W_{gi} (p_i - \bar{p}_g)(\bar{Y}_i - \bar{Y}_g)$ with $\bar{p}_g = \sum_{A_g} W_{gi} p_i$. Thus, if p_i and $\phi(i)$ are constant within A_g , then $\hat{y}_{\text{PS.WR1}}$ will be nearly unbiased even if the cell means $\bar{Y}_i, i \in A_g$, differ. This condition is almost sure to be false as long as one poststratum in a group has a probability of being sparse that is substantially different from the others.

In the case of a common coverage probability in poststratum i , $\phi(i)$, we can also compare the biases of the collapsed cell estimator, \hat{y}_{PS2} , with that of $\hat{y}_{\text{PS.WR1}}$. Using results in the previous paragraph, the bias in (6) can be expressed as

$$\text{bias}(\hat{y}_{\text{PS.WR1}}) = \sum_g W_g \left[\frac{C_{\phi y, g}^* / \bar{\phi}_g + (f_{\max} - 1)C_{p\phi, y, g} / \bar{\phi}_g}{1 + (f_{\max} - 1)(\overline{p\phi})_g / \bar{\phi}_g} \right].$$

Since $1 + (f_{\max} - 1)(\overline{p\phi})_g / \bar{\phi}_g \geq 1$, we can use (5) to obtain

$$\begin{aligned} & |\text{bias}(\hat{y}_{\text{PS.WR1}})| \\ & \leq \left| \sum_g W_g \left[\frac{C_{\phi y, g}^* / \bar{\phi}_g + (f_{\max} - 1)C_{p\phi, y, g} / \bar{\phi}_g}{1 + (f_{\max} - 1)(\overline{p\phi})_g / \bar{\phi}_g} \right] \right| \\ & = |\text{bias}(\hat{y}_{\text{PS2}}) + (f_{\max} - 1) \sum_g W_g C_{p\phi, y, g} / \bar{\phi}_g|. \end{aligned} \quad (7)$$

If $p_i \phi(i)$ and \bar{Y}_i are uncorrelated, the absolute bias of $\hat{y}_{\text{PS.WR1}}$ is less than or equal to that of \hat{y}_{PS2} because $1 + (f_{\max} - 1)(\overline{p\phi})_g / \bar{\phi}_g \geq 1$. When $p_i \phi(i)$ and \bar{Y}_i are correlated, there are two cases to consider: (i) $\text{bias}(\hat{y}_{\text{PS2}}) \geq 0$ and (ii) $\text{bias}(\hat{y}_{\text{PS2}}) < 0$. In the former, the last line of (7) will be less than or equal to the absolute bias of \hat{y}_{PS2} if

$$\frac{-2|\text{bias}(\hat{y}_{\text{PS2}})|}{f_{\max} - 1} \leq \sum_g W_g C_{p\phi, y, g} / \bar{\phi}_g \leq 0.$$

In case (ii), the requirement is

$$0 \leq \sum_g W_g C_{p\phi, y, g} / \bar{\phi}_g \leq \frac{2|\text{bias}(\hat{y}_{\text{PS2}})|}{f_{\max} - 1}.$$

If the covariance between the probability of being sparse and covered, $p_i \phi(i)$, and the cell means, \bar{Y}_i , is small in all groups and the opposite sign of $\text{bias}(\hat{y}_{\text{PS2}})$, then $\hat{y}_{\text{PS.WR1}}$ will be less biased than \hat{y}_{PS2} .

The second alternative is denoted PS.WR2 and is intended to exercise more control over the size of the final weight adjustment than does PS.WR1. In PS.WR1 the final adjustment can be larger than f_{\max} . PS.WR2 seeks to limit the final adjustment to $f_{\max} = 2$ or some other maximum set in advance. The general idea is to first determine which cells should be collapsed together, as was done for PS.WR1. Then weights in the sparse cells are multiplied by f_{\max} . The weights in the non-sparse cell in a collapsed group are then adjusted by a constant factor to bring the estimated population count in the group to the control count. The detailed algorithm for computing weights for PS.WR2 is the following:

- (1) Execute steps (1) - (3) of the algorithm in section 2.3 for PS2.
- (2) In a group containing at least one non-sparse cell, compute the control total in group g as $N_g = \sum_{i \in A_g} N_i$ and the adjusted weight for all units k in $A_{g, sp}$ as $\tilde{w}_k = w_k f_{\max}$.
- (3) Compute the adjusted weight for all units k in $A_{g, sp}$ as $\tilde{w}_k = w_k (N_g - \hat{N}_{g, sp}) / \hat{N}_{g, sp}$ where $\hat{N}_{g, sp} = \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k$ and $\hat{N}_{g, sp} = \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} \tilde{w}_k$.
- (4) The final adjusted weight is then \tilde{w}_k for unit k in group g .

This second weight restricted estimator can be written as $\hat{y}_{\text{PS.WR2}} = (1/N) \sum_g \hat{T}_{g\text{WR2}}$ where

$$\begin{aligned} \hat{T}_{g\text{WR2}} &= \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} f_{\max} w_k y_k \\ &+ \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} \frac{N_g - f_{\max} \hat{N}_{g, sp}}{\hat{N}_{g, sp}} w_k y_k \end{aligned}$$

where $\hat{N}_{g, sp} = \sum_{i \in A_{g, sp}} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k$. The expectation of $\hat{T}_{g\text{WR2}}$ over the coverage, sparseness, and sampling mechanisms is

$$E_c E_{sp} E_\pi (\hat{T}_{g\text{WR2}}) = f_{\max} \tilde{T}_g^c + \frac{N_g - f_{\max} \tilde{N}_g^c}{\tilde{N}_g^c - \tilde{N}_g^c} (T_g^c - \tilde{T}_g^c)$$

where $\tilde{N}_g^c = \sum_{i \in A_g} p_i N_i^c$. After some calculation, the approximate bias of $\hat{y}_{\text{PS.WR2}}$ can be written as

$$\begin{aligned} \text{bias}(\hat{y}_{\text{PS,WR2}}) &= \frac{1}{N} f_{\max} \sum_g \tilde{N}_g^c (\mu_{g,sp}^c - \mu_{g,sp}^c) \\ &+ \sum_g W_g \frac{1}{\sum_{A_g} q_i N_i^c} \\ &\times \left[\sum_{A_g} q_i T_i^c - N_g^{-1} \left(\sum_{A_g} q_i N_i^c \right) \left(\sum_{A_g} T_i \right) \right] \end{aligned}$$

where $\mu_{g,sp}^c = \tilde{T}_g^c / \tilde{N}_g^c$ and $\mu_{g,sp}^c = \sum_{A_g} q_i T_i^c / \sum_{A_g} q_i N_i^c$. Next, note that in the case of a common coverage probability in cell i , $\phi_k = \phi(i)$,

$$\begin{aligned} \sum_{A_g} q_i T_i^c - N_g^{-1} \left(\sum_{A_g} q_i N_i^c \right) \left(\sum_{A_g} T_i \right) \\ = \sum_{A_g} \sum_{h,j \in U_h, k \in U_{j(i)}} (q_i \phi_k - (\bar{q}\phi)_g) (y_k - \bar{Y}_g) \\ = \sum_{A_g} N_i (q_i \phi(i) - (\bar{q}\phi)_g) (\bar{Y}_i - \bar{Y}_g) \\ \equiv N_g (C_{\phi y g}^* - C_{p\phi, y, g}) \end{aligned}$$

with $q_i = 1 - p_i$, $(\bar{q}\phi)_g = \sum q_i \phi_k / N_g$, and $C_{\phi y g}^*, C_{p\phi, y, g}$ were defined previously. Next, use the fact that $\sum_{A_g} q_i N_i^c = N_g - \tilde{N}_g^c$ to define $P_g^c = N_g^c / N_g$, the proportion of units covered in group g , and $\tilde{P}_g^c = \tilde{N}_g^c / N_g$, the expected proportion covered in sparse cells in group g . Then, the bias can also be written as

$$\begin{aligned} \text{bias}(\hat{y}_{\text{PS,WR2}}) &= \frac{1}{N} f_{\max} \sum_g \tilde{N}_g^c (\mu_{g,sp}^c - \mu_{g,sp}^c) \\ &+ \sum_g W_g \frac{C_{\phi y g}^* - C_{p\phi, y, g}}{P_g^c - \tilde{P}_g^c}. \end{aligned} \quad (8)$$

Judging from (8), $\hat{y}_{\text{PS,WR2}}$ will be approximately unbiased if the mean per unit for the units covered by the frame in each collapsed cell is the same in the sparse cells, $\mu_{g,sp}^c$, as in the nonsparse cells, *i.e.*, $\mu_{g,sp}^c = \mu_{g,sp}^c$, and the covariances, $C_{\phi y g}^*$ and $C_{p\phi, y, g}$, are both 0. The latter is accomplished by combining cells with the same means, \bar{Y}_i . Combining cells with equal coverage rates does not result in $\hat{y}_{\text{PS,WR2}}$ being unbiased. This is more restrictive than for $\hat{y}_{\text{PS,WR1}}$, which is unbiased if either the coverage rates or the means are the same in all cells in a collapsed group.

4. An empirical investigation

To test some of the ideas presented earlier, we conducted a simulation study of the bias properties of alternative methods of poststratification. We also examined the performance of one variance estimator that is often used in practice.

4.1 Study population

The population used in the simulation was extracted from the 2003 National Health Interview Survey (NHIS) person

public-use file. A subset of the NHIS was created with 21,664 persons. These were divided into 25 strata with each having six PSUs. The strata and PSU's are based on those in the NHIS public use file, but sets of three strata were collapsed together to create new design strata for the study population. We used four binary variables (0-1 characteristics) for the simulation, each of which is based on a person's self-report:

Health insurance coverage - whether a person was covered by any type of health insurance;

Physical, mental, or emotional limitation - whether a person was limited in any of these ways;

Medical care delayed - whether a person delayed medical care or not because of cost in last 12 months;

Overnight hospital stay - whether a person stayed overnight in a hospital in last 12 months.

Table 2 shows the percentages of persons with these four characteristics in cells formed by age and sex. These 16 (age \times sex) cells are the initial set of poststrata used in estimation. The percentages can vary substantially among the cells, depending on the characteristic. For, example, 18-24 year olds are much more likely to have no health insurance; children under age 5 and the elderly age 65 and over are much more likely to have had a hospital stay. Collapsing cells together that have different means, or proportions in this case, has the potential to introduce bias, as noted earlier.

We also created one artificial binary variable that had a common mean of 0.20 regardless of the unit's poststratum membership. In that case all estimators, including the Hájek estimator, will be unbiased regardless of coverage rates. Also, the conventional thinking that collapsing of cells may reduce variances by smoothing out extreme weight adjustments may hold for this variable.

4.2 Sample design

Two sample PSU's were selected in each stratum with probability proportional to size (PPS) with the size being the count of persons in each PSU. Sampling of PSU's was done with-replacement to simplify variance estimation. If without-replacement sampling had been used, then a more elaborate method of selection and variance estimation would have been needed (see, *e.g.*, Särndal, Swensson, and Wretman 1992, chapter 3). In each sample PSU, 20 persons were selected by simple random sampling without replacement for a total of 1,000 persons in each sample. For each combination of parameters discussed below, 2,000 samples were selected.

Sixteen initial poststrata were used which were the cross of the eight age groups, shown in Table 2, with gender. In

each sample, we computed the estimators of population proportions, described earlier in sections 2-3 - the Hájek estimator, \hat{y}_n , the poststratified estimator \hat{y}_{PS1} , that uses all 16 poststrata, the poststratified estimator with collapsing of cells, \hat{y}_{PS2} , and the two weight-restricted estimators, $\hat{y}_{PS.WR1}$ and $\hat{y}_{PS.WR2}$. The simulation code was written in the R language (R Development Core Team 2005) with extensive use of the R survey package (Lumley 2004, 2005).

4.3 Coverage mechanisms

Five sets of coverage mechanisms, shown in Table 3, were employed to filter the population before the PSU's were sampled. The coverage ratios varied by poststratum and were different for each of the five characteristics for which proportions were estimated. The coverage ratios specific to each of the five characteristics are named C1 through C5 in Table 3. These coverage ratios were artificially created based on the population means for each age and sex group. Poorer coverage was assigned to groups with larger percentages with a characteristic for health insurance coverage and limitations; the opposite was true for delayed medical care and hospital stays. In C5 the coverage ratios are quite variable and are intended to lead to coverage adjustments that vary substantially among the initial set of 16 poststrata. Although the rates in Table 3 are low, they are comparable to or higher than those for BRFSS in Table 1. In applying these rates, we randomly selected a subset of the population to be in the sample frame for each

sample that was selected. For example, if the coverage ratio in the poststratum of males younger than 5 years old is 0.9, then 90% of the population in that poststratum was randomly selected to stay in the sampling frame while the rest had a zero probability of being sampled.

4.4 Collapsing rules

We set up situations where the conditions for unbiasedness in sections 2 and 3 can be violated when cells were collapsed in the simulations. Each of the estimators, \hat{y}_{PS2} , $\hat{y}_{PS.WR1}$, and $\hat{y}_{PS.WR2}$ involve cell collapses. If the IAF (poststratification factor) in an initial poststratum, N_i / \hat{N}_{pi} , exceeds the maximum allowable adjustment, f_{max} , or if the cell sample size is less than a minimum, $n_{i,min}$, we call this poststratum a "sparse" cell and collapse it with a neighboring cell. We used two methods of determining neighbors, designated here as "adjacency" and "close-mean".

In adjacency collapsing, the neighbors of a specific cell are defined as the cells either horizontally or vertically adjacent to it in the age \times sex table. For example, in the following, abbreviated table, the neighbors of cell 3 are the shaded cells 2, 4, and 7.

1	5
2	6
3	7
4	8

Table 2 Percentages of persons with four health-related characteristics in groups formed by age and sex

Age	Population Counts			Not covered by health insurance			Percentage of persons with characteristic Physical, mental, emotional limitations			Delayed medical care in last 12 months			Hospital stay in last 12 months		
	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total	Male	Female	Total
<5	843	795	1,638	10	9	9	4	3	3	3	4	3	17	15	16
5 - 17	2,271	2,082	4,353	13	14	13	10	6	8	4	4	4	2	1	2
18 - 24	998	1,031	2,029	37	31	34	4	4	4	8	11	9	3	14	8
25 - 44	2,971	3,207	6,178	28	23	25	7	7	7	9	10	9	3	10	6
45 - 64	2,421	2,597	5,018	14	14	14	16	19	18	7	11	9	8	10	9
65 - 69	305	384	689	2	1	2	24	29	27	3	8	6	15	14	14
70 - 74	275	344	619	1	1	1	34	32	33	2	5	4	18	15	17
75+	423	717	1,140	1	1	1	41	48	45	2	2	2	22	22	22
Total	10,507	11,157	21,664	18	16	17	12	13	13	6	8	7	7	10	8

Table 3 Coverage ratios used in the simulations

Age	C1: Not covered by health insurance		C2: Physical, mental, emotional limitations		C3: Delayed medical care in the last 12 months		C4: Hospital stay in last 12 months		C5: Common Mean Y	
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
<5	0.9	0.9	0.9	0.9	0.5	0.5	0.8	0.8	0.9	0.8
5 - 17	0.8	0.8	0.9	0.6	0.5	0.5	0.5	0.5	0.7	0.2
18 - 24	0.5	0.5	0.8	0.8	0.8	0.8	0.5	0.8	0.4	0.4
25 - 44	0.5	0.5	0.8	0.8	0.8	0.8	0.5	0.8	0.6	0.5
45 - 64	0.8	0.8	0.5	0.6	0.8	0.8	0.5	0.8	0.3	0.8
65 - 69	0.9	0.9	0.5	0.6	0.5	0.5	0.8	0.8	0.4	0.4
70 - 74	0.9	0.9	0.5	0.5	0.5	0.5	0.8	0.8	0.2	0.7
75+	0.9	0.9	0.5	0.5	0.5	0.5	0.8	0.8	0.8	0.9

In the adjacency method, a sparse cell was collapsed with the neighbor with the smallest poststratification factor. In close-mean collapsing, a sparse cell was collapsed with the nonsparse cell whose unweighted sample mean was closest to that of the sparse cell.

Two different values of f_{\max} were used in the simulations – $f_{\max} = 2$ and 1.8 . Use of $f_{\max} = 1.8$ leads to more collapsing of cells than $f_{\max} = 2$ and exhibits more of the biases (for the characteristics other than the artificial one with a common mean) noted in sections 2 and 3 caused by combining of cells with different means or different coverage rates. The minimum cell size was set to $n_{i,\min} = 25$. Of course, in practice many variations are used to decide which combinations of cells are allowable. We have used just two of the possibilities for illustration in the simulation.

Once all of the sparse cells and their neighbors are identified, the collapsing process proceeds sequentially from cell 1. In a survey with many potential poststrata defined in advance, these procedures might have to be performed iteratively to eliminate all sparse cells. In this simulation, we performed only one round of collapsing.

4.5 Variance estimation

For each of the estimators of a proportion, a linearization variance estimate was calculated. Each of the variance estimators is based on the linear substitute method (e.g., see Särndal *et al.* 1992, chapter 5). The variance estimates for all estimators of proportions were computed using the `svydesign`, `poststratify`, and `svymean` functions in the R survey package. The general, theoretical approach is to make a linear approximation for a particular estimator. The linear approximation is rearranged so that the estimator is written as a sum of weighted PSU totals, and the variance estimator for with-replacement PSU sampling is used. The estimators \hat{y}_{PS1} , \hat{y}_{PS2} , $\hat{y}_{\text{PS.WR1}}$, and $\hat{y}_{\text{PS.WR2}}$ are treated as standard poststratified estimators for the purposes of variance estimation. For \hat{y}_{PS1} , define the following:

$u_k = N_i / \hat{N}_{i\pi} (y_k - \hat{y}_i)$, $k \in s_i$, with $\hat{y}_i = \hat{T}_{i\pi} / \hat{N}_{i\pi}$ and s_i being the set of all sample units in poststratum i ; u_k is known as a linear substitute;

$\tilde{u}_{hj+} = \sum_{i: k \in s_{h(i)}} w_k u_k$; and

$\bar{\tilde{u}}_{h++} = \frac{1}{n_h} \sum_{j \in s_h} \tilde{u}_{hj+}$

The variance estimator for \hat{y}_{PS1} is then

$$v(\hat{y}_{\text{PS1}}) = \frac{1}{N^2} \sum_h \frac{n_h}{n_h - 1} \sum_{j \in s_h} (\tilde{u}_{hj+} - \bar{\tilde{u}}_{h++})^2. \quad (9)$$

For the collapsed stratum estimator, \hat{y}_{PS2} , (9) applies with the linear substitute defined as

$$u_k = \frac{N_g}{\hat{N}_{g\pi}} (y_k - \hat{y}_g), \quad k \in s_g,$$

with $\hat{y}_g = \hat{T}_{g\pi} / \hat{N}_{g\pi}$ and s_g is the set of all sample units in group g .

In the cases of PS.WR1 and PS.WR2, we calculate the final weights as described in section 3 and call the R `poststratify` function. This results in the linear substitute being computed as

$$u_k = \frac{N_g}{\hat{N}_g} (y_k - \hat{y}_g) = y_k - \hat{y}_g$$

because $\hat{N}_g = \sum_{i \in A_g} \sum_{k \in s_i} \tilde{w}_k = N_g$. The mean \hat{y}_g is computed as

$$\begin{aligned} \hat{y}_g &= \sum_{i \in A_g} \sum_{k \in s_i} \tilde{w}_k y_k / \sum_{i \in A_g} \sum_{k \in s_i} \tilde{w}_k \\ &= \sum_{i \in A_g} \sum_{k \in s_i} \tilde{w}_k y_k / N_g. \end{aligned}$$

The weighted linear substitute is then $u_k = \tilde{w}_k (y_k - \hat{y}_g)$ and $\tilde{u}_{hj+} = \sum_{i: k \in s_{h(i)}} \tilde{w}_k u_k$.

In the cases of PS2, PS.WR1, and PS.WR2, these variance estimators do not account for the dynamic nature of cell collapsing which can vary from sample to sample. Consequently, there is a source of variation that is not accounted for, and we can anticipate that the variance estimates will be somewhat too small compared to empirical, simulation variances.

4.6 Simulation results

Tables 4-7 summarize results for coverage correction errors, relative biases of estimated proportions, variances of alternative estimators, and confidence interval coverage using linearization variance estimators. Table 4 shows average absolute coverage correction error, defined as

$$\bar{e} = (DI)^{-1} \sum_{d=1}^D \sum_{i=1}^I |\hat{N}_{di} / N_i - 1| \quad (10)$$

where d is one of the $D=2,000$ samples and \hat{N}_{di} is the estimated number of units in poststratum i based on the final weights for a particular estimator (Hájek, PS2, PS.WR1, or PS.WR2). The value of \bar{e} is 0 for the poststratified estimator with no cell collapsing, PS1, since it corrects coverage error completely in each of the 16 poststrata. To illustrate how the average coverage correction errors can vary, we estimated the proportions for the health insurance and common mean Y variable using the C1 and C5 frame coverage ratios. For most combinations of coverage ratios, collapsing method, and adjustment bound, PS.WR1 more effectively corrects for coverage error than the standard

collapsing estimator, PS2. For example, $\bar{e} = 0.086$ with PS.WR1 for (health insurance, adjacency collapsing, $f_{\max} = 2$) while PS2 has $\bar{e} = 0.120$. In contrast, PS.WR2 is somewhat worse than PS2 in coverage correction.

Table 5 presents the relative biases (relbias), defined as $100 \sum_{d=1}^D (\hat{y}_d - \bar{Y}) / \bar{Y}$ where \hat{y}_d is one of the estimates of proportion for sample d . The Hájek estimates are badly biased for the first four characteristics since they include no correction for the differential undercoverage among the cells. The relbiases range from -12.1% for limitations to 13.4% for hospital stay. As noted in section 2, the bias can be either positive or negative, depending on the correlation of coverage rates and cell means.

Poststratification with no collapsing of cells (PS1) gives nearly unbiased estimates while the alternatives - PS2, PS.WR1, and PS.WR2 - all introduce a bias when using adjacency collapsing for the first four characteristics. The number of poststrata after collapsing, shown in Table 5, ranges from 6 to 16 when $f_{\max} = 2$ and from 5 to 13 when $f_{\max} = 1.8$. The relative biases of PS2, using adjacency collapsing, range from -4.4% to 6.2% when $f_{\max} = 2$ and from -6.5 to 9.4% when $f_{\max} = 1.8$. With adjacency collapsing, The alternatives, PS.WR1 and PS.WR2, have biases that are intermediate between PS1 (no collapsing) and PS2. PS.WR1, in particular, is reasonably competitive with PS1 in terms of bias with adjacency collapsing. In contrast, close-mean collapsing yields PS2, PS.WR1, and PS.WR2 estimates that are essentially unbiased when $f_{\max} = 2$. With mean collapsing and $f_{\max} = 1.8$, PS2 and PS.WR2 are still somewhat biased, but PS.WR1 compares well with PS1. For the fifth characteristic (Common mean Y), all estimators are nearly unbiased, regardless of collapsing method, as expected.

One justification that is conventionally given for collapsing cells is that extreme weights will be reduced and variances of estimates will, in turn, be reduced. Table 6 shows the ratios of the empirical variances of estimated proportions as a proportion of the variance of PS1. The Hájek estimates have variances that are about 12% and 18%

smaller than those of PS1 for health insurance and limitations, but are more variable than PS1 for delayed care and hospital stay. These results also make it clear that the variance of a poststratified estimator can be either increased or decreased by collapsing. There are some minor variance gains from using PS2 for some combinations for the first four variables, but with (adjacency, $f_{\max} = 2$) the PS2 variance of hospital stay is 17% larger than that of PS1. With (adjacency, $f_{\max} = 1.8$), PS2 is 23% more variable for hospital stay. PS.WR1 does not have the extreme variances of PS2 in adjacency collapsing; like PS2, PS.WR2 has larger variance for hospital stay in adjacency collapsing. When close-mean, rather than adjacency, collapsing is used, variances of PS2, PS.WR1, and PS.WR2 are much closer to those of PS1. However, for the Common Mean Y variable, collapsing always reduces variance. The reductions are almost 20% for adjacency collapsing.

The right-hand section of Table 6 lists the ratios of the empirical mean square errors (MSEs) of estimated proportions as a proportion of the MSE of PS1. With a few exceptions, PS2 is the worst choice of the poststratified estimators for the first four characteristics regardless of the combination of variable, f_{\max} , and collapsing method. When $f_{\max} = 1.8$, the choice that leads to more collapsing, the MSEs of PS2 range from 1.8% to 44.2% larger than those of PS1. The MSEs of both PS.WR1 and PS.WR2 are near those of PS1 with the exception of (hospital stay, $f_{\max} = 1.8$, adjacency) where the 6.3% bias of PS.WR2 leads to an MSE 25.6% larger than that of PS1. Close-mean collapsing is preferable to adjacency collapsing, although for the first four characteristics none of the estimators have smaller MSEs than PS1, which does not use collapsing.

The estimators again perform differently for the Common mean Y variable. The MSEs of Hájek, PS2, PS.WR1, and PS.WR2 are all less than that of PS1. The Hájek estimator has the smallest MSE, owing to the fact that poststratification is unnecessary to correct bias in estimating the mean.

Table 4 Average absolute coverage correction error as defined in expression (10)

Collapsing Method	Adjustment Bound f_{\max}	Hájek	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment)
C1 Coverage ratios for health insurance variable					
Adjacency	2	0.257	0.120	0.086	0.221
Close mean	2	0.257	0.080	0.127	0.281
Adjacency	1.8	0.256	0.150	0.085	0.202
Close mean	1.8	0.256	0.101	0.109	0.258
C5 Coverage ratios for common mean Y variable					
Adjacency	2	0.442	0.326	0.196	0.331
Close mean	2	0.441	0.321	0.203	0.370
Adjacency	1.8	0.442	0.330	0.206	0.376
Close mean	1.8	0.442	0.337	0.214	0.446

Table 5 Relative biases (in percent) of estimated proportions. (Figures for Hájek and PS1 are not affected by collapsing and are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Range of no. of poststrata after collapsing	Hájek	PS1 (no collapsing)	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weight adjustment)
Adjacency collapsing, adjustment bound = 2						
Health insurance	(10, 16)	-11.5	0.1	-4.4	1.0	-1.4
Limitations	(8, 15)	-12.1	-0.3	-2.0	0.1	-1.0
Delayed care	(6, 14)	8.2	-0.2	2.2	-0.6	0.9
Hospital stay	(9, 16)	13.4	0.2	6.2	-0.7	2.8
Common mean \bar{Y}	(5, 11)	0.3	0	0.4	0.4	0.6
Close-mean collapsing, adjustment bound = 2						
Health insurance	(10, 16)	-11.5	0.1	-0.5	0.5	-0.3
Limitations	(8, 15)	-12.1	-0.3	-1.2	0.2	-1.1
Delayed care	(6, 14)	8.2	-0.2	-0.3	-0.3	-0.2
Hospital stay	(9, 16)	13.4	0.2	0.4	0.1	0.4
Common mean \bar{Y}	(6, 11)	0.3	0	0.2	0.1	0.2
Adjacency collapsing, adjustment bound = 1.8						
Health insurance	(7, 13)	-11.5	0.1	-6.5	0.7	-3.5
Limitations	(7, 12)	-12.1	-0.3	-3.4	0.3	-2.0
Delayed care	(5, 11)	8.2	-0.2	3.5	-0.4	2.5
Hospital stay	(5, 12)	13.4	0.2	9.4	0.0	6.3
Common mean \bar{Y}	(5, 9)	0.3	0.1	0.5	0.6	0.6
Close-mean collapsing, adjustment bound = 1.8						
Health insurance	(6, 13)	-11.5	0.1	-1.6	0.3	-1.7
Limitations	(7, 12)	-12.1	-0.3	-2.7	0.9	-2.4
Delayed care	(5, 10)	8.2	-0.2	0.2	-0.3	0.5
Hospital stay	(5, 12)	13.4	0.2	1.5	0.3	2.0
Common mean \bar{Y}	(5, 10)	0.3	0.2	0.3	0.3	0.3

Table 6 Ratio of variances (or MSEs) to the variance (or MSE) of the poststratified estimator (PS1) with no collapsing. (Figures for Hájek are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Ratio of variances to the variance of the poststratified estimator (PS1)				Ratio of MSEs to the MSE of the poststratified estimator (PS1)			
	Hájek	PS2	PS.WR1	PS.WR2 (fixed	Hájek	PS2	PS.WR1	PS.WR2 (fixed
		(standard collapsing)	(truncate weights then collapse)	maximum weight adjustment)		(standard collapsing)	(truncate weights then collapse)	maximum weight adjustment)
Adjacency collapsing, adjustment bound = 2								
Health insurance	0.877	1.014	1.025	0.991	1.500	1.101	1.018	1.006
Limitations	0.821	0.966	1.035	0.977	1.555	1.008	1.017	0.992
Delayed care	1.099	1.023	1.003	1.000	1.239	1.023	1.000	1.000
Hospital stay	1.290	1.169	1.000	1.073	1.733	1.244	1.000	1.070
Common mean \bar{Y}	0.755	0.805	0.908	0.818	0.752	0.801	0.904	0.826
Close-mean collapsing, adjustment bound = 2								
Health insurance	0.877	1.013	1.014	1.008	1.500	1.006	1.006	1.006
Limitations	0.821	0.999	1.025	0.994	1.555	1.008	1.017	1.008
Delayed care	1.099	0.997	0.998	1.001	1.239	1.000	1.000	1.000
Hospital stay	1.290	1.011	1.000	1.008	1.733	1.012	1.000	1.012
Common mean \bar{Y}	0.776	0.935	0.974	0.902	0.781	0.933	0.973	0.906
Adjacency collapsing, adjustment bound = 1.8								
Health insurance	0.877	0.960	1.044	0.976	1.500	1.179	1.024	1.048
Limitations	0.821	0.939	1.032	0.961	1.555	1.034	1.017	1.000
Delayed care	1.099	1.051	0.991	1.032	1.239	1.057	0.989	1.034
Hospital stay	1.290	1.225	1.043	1.201	1.733	1.442	1.023	1.256
Common mean \bar{Y}	0.780	0.815	0.882	0.828	0.779	0.816	0.893	0.829
Close-mean collapsing, adjustment bound = 1.8								
Health insurance	0.877	1.010	1.006	1.019	1.500	1.018	1.000	1.024
Limitations	0.821	0.983	1.051	0.975	1.555	1.034	1.034	1.017
Delayed care	1.099	1.003	0.995	1.001	1.239	1.000	1.000	1.000
Hospital stay	1.290	1.052	1.001	1.059	1.733	1.035	1.000	1.047
Common mean \bar{Y}	0.771	0.924	0.958	0.876	0.778	0.932	0.959	0.879

Table 7

Coverage rates in percent of 95% confidence intervals computed using t -distribution with 25 DF. (Figures for Hájek and PS1 are not affected by collapsing and are repeated in the four sections of the table to facilitate comparisons)

Characteristic	Hájek	PS1 (no collapsing)	PS2 (standard collapsing)	PS.WR1 (truncate weights then collapse)	PS.WR2 (fixed maximum weights adjustment)
Adjacency collapsing, adjustment bound = 2					
Health insurance	75.9	93.8	90.1	94.4	93.6
Limitations	70.9	94.5	93.1	94.5	93.9
Delayed care	92.0	94.0	94.5	94.0	94.6
Hospital stay	82.2	94.5	91.8	94.3	93.9
Common mean \bar{Y}	94.8	93.8	94.6	94.4	94.5
Close-mean collapsing, adjustment bound = 2					
Health insurance	75.9	93.8	93.7	94.2	94.0
Limitations	70.9	94.5	93.0	94.3	93.5
Delayed care	92.0	94.0	93.6	93.9	93.8
Hospital stay	82.2	94.5	94.3	94.7	94.6
Common mean \bar{Y}	93.7	92.9	92.2	92.5	93.2
Adjacency collapsing, adjustment bound = 1.8					
Health insurance	75.9	93.8	87.5	94.1	92.4
Limitations	70.9	94.5	92.0	94.2	93.3
Delayed care	92.0	94.0	94.8	94.5	94.5
Hospital stay	82.2	94.5	88.4	94.0	91.1
Common mean \bar{Y}	94.8	94.1	94.3	94.8	94.7
Close-mean collapsing, adjustment bound = 1.8					
Health insurance	75.9	93.8	92.8	94.3	93.3
Limitations	70.9	94.5	92.3	94.5	93.0
Delayed care	92.0	94.0	93.8	94.0	94.4
Hospital stay	82.2	94.5	93.8	94.6	93.8
Common mean \bar{Y}	94.9	94.5	93.6	93.8	94.8

Table 7 reports the empirical coverages of 95% CI's computed using the estimated proportions and the linearization variance estimator that naturally accompanies each. A t -distribution with 25 degrees of freedom is used in all cases. The Hájek coverage rates are extremely poor, as expected, ranging from 70.9% to 92% for the first four characteristics. The poststratified estimators, PS1 and PS.WR1 provide 93.8% to 94.7% coverage, *i.e.*, near the nominal 95%. In contrast, PS2 coverage is somewhat poor for Health insurance and hospitalization, especially for (adjacency, $f_{\max} = 1.8$) where the coverages are 87.5% and 88.4%. Coverage rates for PS.WR2 are slightly less than for PS.WR1 but are reasonably close to nominal. Use of close-mean collapsing generally improves the cases of poor coverage found with adjacency. For Common Mean \bar{Y} coverages are good, ranging from 92.2% to 94.9%.

In summary, the weight-restricted estimators, PS.WR1 and PS.WR2, have some advantage over the standard collapsing estimator, PS2. They are generally less biased and retain more of the undercoverage adjustment than does PS2. However, the most critical element in bias-control is how the cells are collapsed in the first place. Collapsing using nearness of cell means or coverage rates is far more preferable than collapsing using some adjacency criterion based on neither of these. Only when cell means were equal did we observe any gain in MSE from collapsing cells. However, equality of cell means is the exception in practice.

5. Concluding remarks

Designers of surveys of households or establishments often have a lengthy list of poststrata or cells in mind when they develop weighting systems. If the sample size in a poststratum is small or the sample estimate of the population count in a poststratum is much different from an external control count, the poststratum may be collapsed with an adjacent one. The conventional justification for collapsing is that the possibility of creating extreme weights is reduced as are variances of estimates.

However, a poor choice of the method for collapsing has at least two undesirable consequences: (i) deficient frame or sample coverage in some cells is not completely corrected and (ii) estimates from the standard approach to collapsing may be quite biased. The latter problem can result in confidence intervals that cover at much less than the nominal rate. Collapsing leads to bias when coverage rates, cell means, or both are correlated within a collapsed poststratum. The bias can be either positive or negative, depending on the correlation.

Cells should be collapsed based on similarity of coverage rates, population cell means, or both in order to avoid bias. This method of collapsing can be much different from standard procedures that only collapse "adjacent" cells, *e.g.*, by combining contiguous age groups. If the adjacency coincides with cells that have similar coverage rates or

means, no bias results. But, this should be checked rather than assumed.

There are at least two practical issues with collapsing based on cell means. One is that, while the theory directs us to collapse based on population means, in a particular sample we will only have estimates for the population covered by the frame. Coverage may be so deficient that the means of the covered and non-covered parts of the population are substantially different, even within the initial poststrata. This would be a case of "nonignorable non-coverage." If so, poststratification based only on the initial set of cells or combinations of them cannot correct coverage bias. A second practical issue is that data on many items are collected in most surveys. Collapsing based on the cell means for one variable may not work well for other variables. In that case, the compromise, suggested by Little and Vartivarian (2005) for nonresponse adjustment, of collapsing based on some weighted average of the means of an important set of variables should be a good solution.

Extensions of this research would be to examine the performance of the class of calibration estimators in correcting coverage errors. Poststratification is a special case. When categories of qualitative auxiliaries are combined due to small sample sizes or other reasons, the same bias problems we have illustrated here may be introduced in more general calibration estimators. One method of allowing some flexibility to depart from controls while retaining important auxiliaries is already available in Rao and Singh (1997). The effect of their proposals on coverage bias needs to be investigated.

Acknowledgements

The authors are indebted to the Associate Editor and the referees for their careful reviews and constructive comments. This paper reports the general results of research undertaken in part by National Center for Health Statistics (NCHS) staff. The views expressed are attributable to the authors and do not necessarily reflect those of the NCHS. The work of R. Valliant was partially supported by Professional Services Contracts 200-2004-M-09302 and 200-2006-M-17916 between the University of Michigan and the National Center for Health Statistics.

References

- Bureau of the Census (2002). *Sources and Accuracy of Estimates for Poverty in the United States: 2002*. Washington DC.
- Eltinge, J., and Yansaneh, I. (1997). Diagnostics for formation of nonresponse adjustment cells with an application to income non-response in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 37-45.
- Fuller, W. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- Gonzalez, J.F., Town, M. and Kim, J. (2005). Mean square error analysis of health estimates from the Behavioral Risk Factor Surveillance System for counties along the United States/Mexico border region. *Proceedings of the Section on Survey Methods Research*. Alexandria VA: American Statistical Association.
- Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- Kalton, G., and Maligalig, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 409-428.
- Kim, J.J. (2004). Effect of collapsing rows/columns of weighting matrix on weights. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.
- Kim, J.J., Thompson, J.H., Woltman, H.F. and Vajs, S.M. (1982). Empirical results from the 1980 Census Sample Estimation Study. *Proceedings of Section on Survey Research Methods*, American Statistical Association, 170-175.
- Kim, J.J., and Tompkins, L. (2007). Comparisons of current and alternative collapsing approaches for improved health estimates. Paper presented at the 11th Biennial CDC/ASTDR Symposium on *Statistical Methods*, in Atlanta, Georgia, April 17-18, 2007.
- Kim, J.J., Tompkins, L., Li, J. and Valliant, R. (2005). A simulation study of cell collapsing in poststratification. *Proceedings of the Section on Survey Methods Research*, American Statistical Association.
- Kostanich, D., and Diplo, C. (2000). *Current Population Survey: Design and Methodology*. Technical paper 63. Washington DC: Department of Commerce.
- Lazzeroni, L., and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 161-168.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 1-19.
- Lumley, T. (2005). *Survey: Analysis of complex survey samples*. R package version 3.0-1. University of Washington: Seattle.
- R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, <http://www.R-project.org>.
- Rao, J.N.K., and Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Tompkins, L., and Kim, J.J. (2006). Evaluation of collapsing criteria in sample weighting. Internal NCHS memorandum.
- Tremblay, V. (1986). Practical criteria for definition of weighting classes. *Survey Methodology*, 12, 85-97.

A single frame multiplicity estimator for multiple frame surveys

Fulvia Mecatti¹

Abstract

Multiple Frame Surveys were originally proposed to foster cost savings on the basis of an *optimality* approach. As surveys on *special*, *rare* and *difficult-to-sample* populations are becoming more prominent, a single list of population units to be used as a sampling frame is often unavailable in sampling practice. In recent literature multiple frame designs have been put forward in order to increase population coverage, to improve response rates and to capture differences and subgroups. Alternative approaches to multiple frame estimation have appeared, all of them relying upon the virtual partition of the set of the available overlapping frames into disjointed domains. Hence the correct classification of sampled units into the domains is required for practical applications. In this paper a multiple frame estimator is proposed using a *multiplicity* approach. Multiplicity estimators require less information about unit domain membership hence they are insensitive to misclassification. Moreover the proposed estimator is analytically simple so that it is easy to implement and its exact variance is given. Empirical results from an extensive simulation study comparing the multiplicity estimator with major competitors are also provided.

Key Words: Difficult-to-Sample populations; Dual frame survey; Misclassification; Raking ratio; Variance estimation.

1. Introduction

In classic finite population sampling a basic hypothesis is the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In some cases a set of two or more lists is available for survey purposes. The general case of $Q \geq 2$ lists, singularly partial and possibly overlapping, is known as *Multiple Frame Survey*. Multiple frame surveys were originally introduced (Hartley 1974) as a device for reducing survey costs by achieving the same precision as a customary unique-frame survey. In modern sampling practice, as surveys of *special*, *rare* and *difficult-to-sample* populations are becoming more common (Kalton and Anderson 1986; Sudman and Kalton 1986; Sudman, Sirken and Cowan 1988) it is often the case that a unique list of units does not exist and the population size N is an unknown parameter to be estimated. Recent literature considers multiple frame surveys with the main aim of increasing population coverage, of improving response rates and of capturing differences and subgroups more accurately (Iachan and Dennis 1993; Carlson and Hall 1994; Haines and Pollock 1998; Eurostat 2000). In a recent paper Lohr and Rao (2006) stated: "As the U.S., Canada, and other nations grow in diversity, different sampling frames may better capture subgroups of the population. [...] We anticipate that modular sampling designs using multiple frames will be widely used in the future". A contemporary application could be found in web surveys: the population coverage can be improved and the bias due to the features of the site used for data collection can be reduced by using two or more independent web sites simultaneously. Since the

same unit can visit more than one site involved in the survey, the sites overlap configuring a multiple frame framework.

Estimation in multiple frame surveys, as first developed by Hartley (1962, 1974), is based on the virtual partition of the population (*i.e.*, the unknown union of the Q overlapping frame) into $2^Q - 1$ disjointed *domains* (*i.e.*, the mutually exclusive intersections of frames). Hence the total Y of a study variable y , taken as the parameter to be estimated, is expressed as a sum of *domain totals*. Sample data from the Q frames are used to produce estimates for the domain totals. Estimated domain totals are finally combined to provide estimation for the population total Y . A number of estimators have been developed according to alternative approaches to multiple frame estimation (see Section 2). Since all estimators appearing in literature rely on the partition into the domains as mentioned above, the correct identification of the domain membership of each sampled unit is required for their practical application. This is a strong assumption that may not always be true in practice, as argued for instance in Lohr and Rao (2006). Indeed this implies that every sampled unit should be questioned on both the survey value and on its membership to each frame involved in the survey, in order to be able to correctly classify them into the domains. In addition to the natural risk of misclassification there might also be a risk connected with *confidentiality* and with the *sensitivity* of units to the frame membership which could both increase the rate of non-response and affect the estimator precision. This situation could apply, for instance, when surveying *sensitive* characteristics (private behaviours, addictions ...) or when sampling *elusive* populations (illegal immigrants,

1. Fulvia Mecatti, Department of Statistics, University of Milan-Bicocca, Via Bicocca degli Arcimboldi, 8. Ed. U7, 20126 Milan, Italy. E-mail: fulvia.mecatti@unimib.it.

ex-prisoners, patients...). In the present paper a different approach to estimation in multiple frame surveys is adopted. The concept of unit *multiplicity*, corresponding to the *number* of frames to which units belong, is proposed in alternative to the existing approaches based on the domain membership, *i.e.*, to which frames units belong. An unbiased estimator, naturally insensitive to domain misclassification and applying to any number of frames, is presented. The proposed multiplicity estimator has a simple analytical structure so that it can be easily implemented, while its exact variance is given in a closed form and hence readily estimated for any sample size.

In Section 2 an overall discussion of the main contributions to multiple frame estimation is presented in a unified view and the necessary notation is introduced. In Section 3 a multiplicity estimator is proposed and variance estimation is analysed. An extensive simulation study comparing the proposed estimator with major competitors is presented in Section 4.

2. Optimum, pseudo-optimum and single frame estimation

Although literature has mostly dealt with the dual frame case ($Q = 2$), a general theoretical framework for multiple frame surveys ($Q \geq 2$) has been recently provided in Lohr and Rao (2006). By using their multiple frame notation, different estimation approaches are briefly reviewed and the available estimators are presented in a unified way which highlights their dependency on the domain membership of the sampled units.

Let $A_1 \cdots A_q \cdots A_Q$ be a collection of $Q \geq 2$ overlapping frames, the union of which offers a population coverage adequate for survey objectives. Let the index sets K be the subsets of the range of the frame index $q = 1 \cdots Q$. For every index set $K \subseteq \{1 \cdots q \cdots Q\}$ a domain is defined as the set $D_K = (\cap_{q \in K} A_q) \cap (\cap_{q \notin K} A_q^C)$, where C denotes complementation.

Let the *domain membership indicator* be the indicator $\delta_i(K)$ taking value 1 if unit i is included in domain D_K and 0 otherwise. The estimating total Y over the (unknown) union of the Q overlapping frames is then expressed as a sum over the set of $2^Q - 1$ disjointed domains

$$Y = \sum_{i \in \bigcup_q A_q} y_i = \sum_K \sum_{i \in \bigcup_q A_q} \delta_i(K) y_i. \quad (1)$$

Let s_q be a sample selected from frame A_q under a given design, independently for $q = 1 \cdots Q$. A general expression of a multiple frame estimator based on the domain classification is then

$$\hat{Y} = \sum_K \sum_{q \in K} \sum_{i \in s_q} w_i^{(q)} \delta_i(K) y_i. \quad (2)$$

Note that when an unbiased estimator for the total Y is given, an estimator for the population size N is also given by simply substituting sample values y_i by 1's.

Estimators available in literature result from setting weights $w_i^{(q)}$ in (2) according to three main approaches. Since multiple frame surveys were originally put forward with the aim of fostering cost savings by achieving equal or greater precision than a customary unique-frame survey, an *optimum* approach was first suggested by using optimum weights $w_{i, \text{opt}}^{(q)}$ in (2), *i.e.*, by minimizing the estimator variance (Hartley 1962, 1974; Lund 1968; Fuller and Burmeister 1972). Optimum estimators have optimal theoretical properties (Skinner 1991; Lohr and Rao 2000) but present practical problems due mainly to their complexity (explicit though complex formulae for optimum weights $w_{i, \text{opt}}^{(q)}$ with any number of frames are given in Lohr and Rao 2006, Section 3). Moreover, optimum weights depend on unknown population covariances so that they must be estimated from sample data. This is both computationally complex and affects optimality since the extra variability in estimating the covariances leads to larger mean square errors (Lohr and Rao 2006, Section 7).

In order to improve the applicability, a *single frame* (SF) approach has been proposed by using *fixed weights* which ensure design-unbiasedness. For simple random sampling in every frame, the SF estimator is given by substituting weights $w_i^{(q)}$ in (2) with $w_{i, \text{SF}}^{(q)} = w^{(K)} = (\sum_{q \in K} f_q)^{-1}$ where $f_q = n_q / N_q$ denotes the frame sampling fraction (Bankier 1986; Kalton and Anderson 1986; Skinner 1991; Skinner, Holmes and Holt 1994). Since fixed weights usually differ from optimum weights, the SF estimator is generally less efficient than an optimum estimator (Lohr and Rao 2000). Finally a *pseudo-optimum approach* was proposed (Skinner and Rao 1996; Lohr and Rao 2000) in order to achieve both a wider applicability than optimum estimators and to improve efficiency compared with the SF approach. A pseudo-maximum likelihood (PML) estimator for multiple frame surveys is given by substituting in (2): $w_{i, \text{PML}}^{(q)} = w^{(K)} = \hat{N}_K / \sum_{q \in K} \sum_{i \in s_q} \delta_i(K) = \hat{N}_K / n_K$ where the estimated domain sizes \hat{N}_K are the solution of a system of non linear equations. Although complex to implement for practical applications (an iterative linear approximation of \hat{N}_K under simple random sampling is given in Lohr and Rao 2006, Section 4.1) the PML estimator retains good theoretical properties from the optimum approach.

Note that formula (2) involves the domain membership indicator $\delta_i(K)$; hence optimum, pseudo-optimum and SF estimators apply only if the correct classification of sample data into the $2^Q - 1$ domains is accomplished.

In the next Section a multiple frame estimator is presented on the basis of a single frame *multiplicity* approach which does not require domain classification.

3. The single frame multiplicity estimator

The notion of multiplicity was first introduced in connection with Network Sampling (Casady and Sirken 1980; Sirken 2004). It is also a tool of the Generalized Weight Share Method (Lavallée 2002; 2007) as well as of the Center Sampling estimation theory (Mecatti 2004) since center sampling and multiple frame surveys are equivalent under certain conditions. In Lohr and Rao (2006), the multiplicity of domain D_K is defined as the cardinality of the index set K . Since domains are mutually exclusive, multiplicity is also a characteristic of every population unit, being the *number* of frames in which each unit is included among the Q involved in the survey.

Let m_i be the multiplicity of unit i . Note that unit multiplicity may be collected simply by asking sampled units *how many frames they belong to*.

Since clearly $\sum_q \sum_{i \in A_q} y_i = \sum_{i \in \bigcup_q A_q} m_i y_i$, it follows that

$$Y = \sum_{q=1}^Q \sum_{i \in A_q} y_i m_i^{-1}. \quad (3)$$

Notice that expression (3), which involves exclusively sums over the frames, represents a practical advantage with respect to equation (1). In fact the domains provide a virtual (unknown) partition of the population while the sample selection is actually performed in the Q overlapping frames. This leads to a SF multiplicity estimator as given by

$$\hat{Y}_M = \sum_{q=1}^Q \sum_{i \in s_q} w_i^{(q)} y_i m_i^{-1} \quad (4)$$

with fixed weights $w_i^{(q)}$ ensuring, for instance, design-unbiasedness. For simple random sampling of every frame we have $w_i^{(q)} = f_q^{-1}$, $\forall i \in s_q$.

Unlike the optimum, PML and SF estimators discussed in Section 2, estimator (4) does not involve the sample membership indicator and it is very simple to implement in practical applications. Furthermore, it is to be noted that for simple random sampling of every frame, the sampled values in multiplicity estimator (4) are weighted by $(f_q m_i)^{-1}$, i.e., by a *specific* frame coefficient; vice versa, in the SF estimator sampled values are weighted by $w_i^{(K)} = (\sum_{q \in K} f_q)^{-1}$, i.e., by an *average* coefficient over the frames involved in each domain. As a consequence \hat{Y}_M is expected to be more accurate than the SF estimator, as confirmed by simulation results. Moreover, owing to its Horvitz-Thompson structure, the exact variance of \hat{Y}_M can be derived in closed form. For

simple random sampling of every frame the estimator variance is given by

$$V(\hat{Y}_M) =$$

$$\sum_{q=1}^Q \frac{N_q(N_q - n_q)}{n_q^2(N_q - 1)} \left[N_q \sum_{i \in A_q} y_i^2 m_i^{-2} - \left(\sum_{i \in A_q} y_i m_i^{-1} \right)^2 \right]. \quad (5)$$

An unbiased variance estimator for simple random sampling of every frame is then

$$\hat{v}(\hat{Y}_M) =$$

$$\sum_{q=1}^Q \frac{N_q(N_q - n_q)}{n_q^2(N_q - 1)} \left[N_q \sum_{i \in s_q} y_i^2 m_i^{-2} - f_q^{-1} \left(\sum_{i \in s_q} y_i m_i^{-1} \right)^2 \right]. \quad (6)$$

The performance of the multiplicity estimator for finite sample sizes has been empirically studied under simple random sampling and compared to major competitors in a simulation study.

4. Simulation study

Several simulation results concerning dual frame estimators have appeared in literature (Bankier 1986; Skinner and Rao 1996; Lohr and Rao 2000). In the general case of $Q \geq 2$ frames, Lohr and Rao (2006) extensively investigated the empirical mean squared errors of a set of eight estimators under optimum, pseudo-optimum and single frame approaches, in a three-frame framework under a two-stage design. Their results suggest that optimum estimators are theoretically optimal but in practice the extra variability in estimating optimum weights leads to larger mean squared errors. Hence the PML estimator appears as the best performer in terms of empirical relative efficiency. Furthermore, their study regarded a case of about 10% of sampled units misclassified into the domains and more research on the effects of misclassification on the estimator performances is recommended.

In the present study pseudo-optimum and single frame estimators are compared with the multiplicity estimator (4), with three main objectives:

- i) to investigate empirical conditions in which the multiplicity estimator results more efficient than the SF estimator (Section 4.2);
- ii) to consider the raking ratio correction to known frame sizes N_q as already proposed in order to improve efficiency of the SF estimator (Section 4.3);
- iii) to explore the effects of increasing rates of misclassification upon the empirical properties of the PML and SF estimators (simple and raked)

versus the natural insensitivity of the multiplicity estimator (Section 4.4).

4.1 Implementation

The simulation study was performed in an artificial three-frame setup and implemented as follows. N population pseudo-values y_i are generated from a Gamma distribution. Some preliminary simulations indicated that both increasing values of the population size N and different values for the Gamma parameters (leading to an asymmetrical and almost symmetrical shape) do not produce significant differences in the pattern of the relative performance of the estimators considered. The study was then conducted by setting $N = 1,200$ and by generating from a Gamma distribution with parameters of 1.5 and 2. Every pseudo-value y_i is randomly assigned to the $Q = 3$ frames according to 3 independent Bernoulli trials with probability $\alpha_q = N_q / N$, $q = 1, 2, 3$. Different scenarios regarding both frame coverage and frame overlapping result from different choices for α_q , under the two constraints: a) $\sum_q \alpha_q \geq 1$ in order to ensure that the 3 frames cover the entire population and b) the 3 frames are non-empty. In some cases, the desired frame overlapping was produced by fixing the ratio N_K / N of the population units included in each domain.

Chosen a set of sampling fractions $f_q = n_q / N_q$, $q = 1, 2, 3$, a simple random sampling is selected independently from every frame, iteratively for 10,000 simulation runs. For a given estimator, say \hat{Y} , the collection of values $\{\hat{Y}_p, p = 1, \dots, 10,000\}$ is assumed as its monte carlo distribution and the empirical mean $E_{mc}(\hat{Y}) = \sum_p \hat{Y}_p / 10,000$ and the empirical mean squared error $MSE_{mc}(\hat{Y}) = \sum_p [\hat{Y}_p - Y]^2 / 10,000$ are calculated. The monte carlo error is controlled by only accepting simulations giving empirical relative bias $RB_{mc}(\hat{Y}) = 100 \cdot |E_{mc}(\hat{Y}) - Y| / Y$ less than 1.5% for those estimators known to be unbiased. Furthermore, by using the exact variance of the multiplicity estimator as given by (6), simulations ensure $|MSE_{mc}(\hat{Y}_M) - V(\hat{Y}_M)| \leq 0.03$. Several different scenarios have been investigated by combining different levels of frame coverage, of frame overlapping and of sampling disproportion, leading to 29 simulated populations. In Figure 1 the simulated populations are represented as points in the plane formed by the two main simulation parameters, namely the (total) frame coverage on the horizontal axis (as given by $\sum_q \alpha_q$) and the sampling disproportion on the vertical axis, i.e., the dispersion among the sampling fractions f_q as measured by $\sum_q \alpha_q |f_q - f_q| / 3^2$. The different shape of populations/ points in Figure 1 indicates different levels of overlapping, namely the total rate of population units classified into the four overlapping domains.

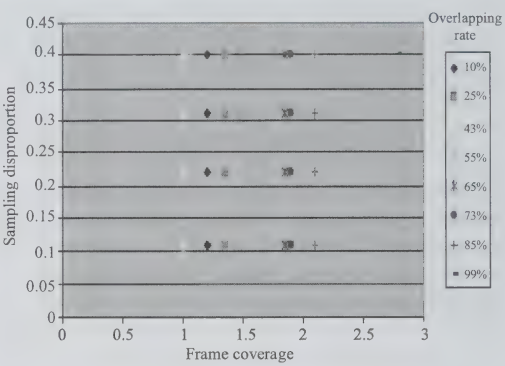


Figure 1 Simulated populations

4.2 Multiplicity versus simple single frame estimation

As noted in Section 3, the multiplicity estimator involves specific frame weights whereas the SF estimator is based on average coefficients. As a consequence the two estimators coincide for constant sample fraction $f_q = f$ in every frame, i.e., for proportionate sampling, and they offer different estimates for disproportionate sampling. Simulation results provide empirical evidence that the multiplicity estimator is more accurate than the simple SF estimator. Estimator \hat{Y}_M is shown to be more efficient in all the cases explored except in one extreme case in which the three frames are almost complete and hence the total overlapping is close to 100%. Neglecting this single case, efficiency gains of \hat{Y}_M over the SF estimator, as measured by a customary empirical efficiency ratio (see Table 1), range from 5% to 48%, and are never less than 26% in half of the simulations. Efficiency of the multiplicity estimator over the SF estimator increases as the sampling disproportion increases (see Table 2) whereas it has resulted as being essentially independent with respect to increasing levels of frame coverage and overlapping.

Table 1
Empirical efficiency ratio of \hat{Y}_M versus SF estimator: Elementary statistics over 28 simulated populations

average	Max	min	median	75 th quantile
0.7425	0.95	0.52	0.74	0.89

Table 2
Empirical efficiency ratio of \hat{Y}_M versus the SF estimator for increasing levels of sampling disproportion

Sampling Disproportion	0.11	0.22	0.31	0.40
Empirical efficiency ratio averaged for different levels of frames coverage/overlapping	0.92	0.81	0.68	0.57

4.3 Raking ratio adjustment

It has been suggested that the raking ratio adjustment using the known frame sizes N_q (Bankier 1986) be used in order to improve efficiency of the simple SF estimator. Theoretical and empirical results that have already appeared in literature confirm that the raking ratio SF estimator (SFrak) can be considerably more efficient than the simple SF estimator (Skinner 1991; Lohr and Rao 2000, 2006; Mecatti 2005).

In order to adjust the multiplicity estimator via raking ratio, knowledge of the domain membership of sampled units has to be assumed. By using this additional, though redundant, information \hat{Y}_M may be rewritten as

$$\hat{Y}_M = \sum_K \sum_{q \in K} (|K| f_q)^{-1} \sum_{i \in s_q} \delta_i(K) y_i \tag{7}$$

where $|K|$ indicates the number of frames involved in domain D_K and it equals unit multiplicity m_i for all $i \in D_K$. Setting the initial weights at $h_{Kq}^{(0)} = (|K| f_q)^{-1}$, the t^{th} iteration of the raking ratio multiplicity estimator (Mrak) is obtained by substituting the following raked weights in (7)

$$h_{Kq}^{(t)} = \begin{cases} \frac{N_q h_{Kq}^{(t-1)}}{\sum_{q \in K} h_{Kq}^{(t-1)} n_q} & \text{if } q \in K \\ h_{Kq}^{(t-1)} & \text{if } q \notin K \end{cases}$$

where $q = Q$ if t is a multiple of Q otherwise $q = t \bmod(Q)$, for $t = 1, 2, \dots$ until convergence.

Simulations regarded different levels of frames coverage combined with different sets of sampling fractions, leading to increasing sampling disproportion.

Empirical results show that Mrak is more efficient than SFrak in 38% of cases explored and it is equally or less efficient in the remaining cases. Efficiency gains range from 3% to 74% and occur for low levels of frame coverage. For increasing frame coverage (and hence increasing overlapping) Mrak estimator is superior to SFrak estimator for high sampling disproportion only. In the other cases, namely for increasing frame coverage/overlapping combined with low to medium sampling disproportion, Mrak can be considerably less efficient than SFrak (see Table 3 for the ten indicative cases) and also severely biased. Thus empirical results suggest that the raking ratio adjustment has better effects under a single frame approach than under a multiplicity approach, although there are conditions in which the latter is still superior. With this respect more research is needed. Particularly, since the raking ratio procedure is in fact a special case of calibration (Deville and Särndal 1992; Deville, Särndal and Sautory 1993), potential improvements might follow by applying the more general calibration to estimator \hat{Y}_{M^*} . Calibration of the multiplicity

estimator, as viewed as a particular case of the Generalized Weight Share Method, is outlined in Lavallée (2002, 2007).

Table 3 Efficiency of Mrak versus SFrak: Ten indicative simulation runs

Frame coverage $\alpha_q = N_q/N$			Sampling fractions $f_q = n_q/N_q$			Empirical efficiency ratio Mrak versus SFrak
0.60	0.60	0.60	0.01	0.95	0.15	0.26
0.35	0.35	0.35	0.80	0.20	0.50	0.54
0.85	0.85	0.85	0.01	0.95	0.15	0.71
0.35	0.40	0.50	0.70	0.80	0.60	0.96
0.85	0.85	0.85	0.80	0.20	0.50	1.01
0.60	0.60	0.60	0.70	0.80	0.60	1.09
0.80	0.50	0.35	0.01	0.95	0.15	1.22
0.35	0.40	0.50	0.01	0.95	0.15	1.63
0.70	0.05	0.95	0.70	0.80	0.60	2.09
0.70	0.05	0.95	0.80	0.20	0.50	5.79

4.4 Misclassification

The aim of the final part of the simulation study is to investigate the sensitivity of the pseudo-optimum (PML) and single frame estimators (simple and raked) to increasing levels of misclassification of sampled units into the domains, with respect to the structural insensitivity of the proposed multiplicity estimator. For a chosen rate of misclassification, the desired number of sampled units to be inexactly classified is taken from the domain with the largest size and randomly assigned to the remaining domains, independently for each frame.

Tables 4 and 5 show elementary statistics summarizing simulation results in the case of exact classification and in the case of slight misclassification equal to 1% of sampled units. Note that for exact classification all the estimators appear unbiased (or nearly unbiased). As regards efficiency, according to other simulation results (Lohr and Rao 2006) SFrak and PML estimators show similar performances. As expected, for exact classification they are more efficient than \hat{Y}_M in all the cases explored (except for two isolated cases) as a consequence of the different amount of information used in the estimation process. However the SF (simple and raked) and PML estimators tends to become biased and less efficient than \hat{Y}_M in presence of just a small amount of misclassification.

Table 4 Relative bias in case of 1% of misclassification: Elementary statistics over the 29 simulated populations

(absolute) RB_{mc} 1% of sampled units misclassified	Average	Min	Max	Median	75 th quantile
\hat{Y}_M	0	0	0	0	0
SF	2.5880	0.83	7.02	2.65	2.13
SFrak	1.7632	0.73	4	1.97	1.65
PML	2.7352	0.23	4.67	3.46	2.87

Table 5

Empirical efficiency ratio of \hat{Y}_M versus SF and PML estimators: Elementary statistics over the 29 simulated populations for exact classification and for slight misclassification

Empirical efficiency ratio	Average	Min	Max	Median	75 th quantile
<i>Exact classification</i>					
SFrak	1.43	0.69	3.21	1.51	1.28
PML	1.41	0.72	3.30	1.47	1.25
<i>1% misclassification</i>					
SF	0.39	0.13	0.71	0.54	0.34
SFrak	0.78	0.13	1.98	0.95	0.74
PML	0.77	0.14	1.94	0.98	0.70

Finally we focused on the case of maximum efficiency of SF, SFrak and PML over \hat{Y}_M for exact classification, namely the case of high frame overlapping/coverage and low sampling disproportion. In this set up, increasing rates of misclassification of sampled units into domains (from 0 to 50%) were investigated. Table 6 and 7 show respectively the relative bias and the efficiency ratio of \hat{Y}_M versus SF, SFrak and PML estimators, for increasing levels of misclassification. It is to be noticed that although the negative effects of misclassification are rapid and severe for all the competitors, the PML estimator emerges as the least affected.

As a conclusion the proposed multiplicity estimator, besides being simple, is recommended when the risk of (even slight) misclassification of sampled units into the domains is a concrete possibility.

Table 6 (absolute) Relative bias for increasing rate of misclassification

% misclassification	\hat{Y}_M	SF	SFrak	PML
0	0	0	0	≈ 0
1%	0	2.57	1.38	4.3
5%	0	13.57	7.15	2.75
10%	0	17.80	14.14	4.56
20%	0	25	25	6
50%	0	144	68	39

Table 7

Empirical efficiency ratio of \hat{Y}_M versus SF, SFrak and PML estimators for increasing rate of misclassification

% misclassification	\hat{Y}_M versus SF	\hat{Y}_M versus SFrak	\hat{Y}_M versus PML
0	0.640	3.210	3.300
1%	0.260	1.040	1.100
5%	0.020	0.060	0.370
10%	0.010	0.020	0.150
20%	0.004	0.004	0.080
50%	≈ 0	0.001	0.006

Acknowledgements

This work was partially supported by a grant from the Italian Ministry of University and Research. The author wishes to thank Jon N.K. Rao for the useful discussion and resourceful advice. Thanks are also due to the editor, two associate editors and two anonymous referees for their constructive comments and suggestions.

References

- Bankier, M.D. (1986). Estimators based in several stratified samples with applications to multiple frame surveys. *Journal of American Statistical Association*, 81, 1074-1079.
- Casady, R.J., and Sirken, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 601-605.
- Carlson, B.L., and Hall, J.W. (1994). Weighting sample data when multiple sample frames are used. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 882-887.
- Deville, J.C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of American Statistical Association*, 87, 376-382.
- Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of American Statistical Association*, 88, 1013-1020.
- Eurostat (2000). *Push and Pull Factors of International Migration*. Country Report-Italy, 3/2000/E/n.5, Bruxelles: European Communities Printing Office.
- Fuller, W.A., and Burmeister, L.F. (1972). Estimators of samples selected from two overlapping frames. *Proceedings of the Social Statistics Sections*, American Statistical Association, 245-249.
- Haines, D.E., and Pollock, K.H. (1998). Combining multiple frame to estimate population size and totals. *Survey Methodology*, 24, 79-88.
- Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Sections*, American Statistical Association, 203-206.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, C, 36, 99-118.
- Iachan, R., and Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of Royal Statistical Society, A*, 149, 65-82.
- Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Editions de l'Université de Bruxelles, Editions Ellipses.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of American Statistical Association*, 95, 271-280.

- Lohr, S., and Rao, J.N.K. (2006). Multiple frame surveys: Point estimation and inference. *Journal of American Statistical Association*, 101, 1019-1030.
- Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Sections, American Statistical Association*, 282-288.
- Mecatti, F. (2004). Center sampling: A strategy for surveying difficult-to-sample populations. *Proceedings of the Statistics Canada Symposium*.
- Mecatti, F. (2005). Single frame estimation in multiple frame survey. *Proceedings of the Statistics Canada Symposium*.
- Sirken, M.G. (2004). Network sample surveys of rare and elusive populations: A historical review. *Proceedings of Statistics Canada Symposium, Keynote Address*.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimator for multiple frame surveys. *Journal of American Statistical Association*, 86, 779-784.
- Skinner, C.J., Holmes, D.J. and Holt, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical review*, 62, 333-347.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of American Statistical Association*, 91, 349-356.
- Sudman, S., and Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.
- Sudman, S., Sirken, M.G. and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 2, 991-996.

Variance estimation for a ratio in the presence of imputed data

David Haziza¹

Abstract

In this paper, we study the problem of variance estimation for a ratio of two totals when marginal random hot deck imputation has been used to fill in missing data. We consider two approaches to inference. In the first approach, the validity of an imputation model is required. In the second approach, the validity of an imputation model is not required but response probabilities need to be estimated, in which case the validity of a nonresponse model is required. We derive variance estimators under two distinct frameworks: the customary two-phase framework and the reverse framework.

Key Words: Imputation model; Nonresponse model; Marginal random hot deck imputation; Reverse framework; Two-phase framework; Variance estimation.

1. Introduction

Variance estimation in the presence of imputed data for simple univariate parameter such as population totals and population means has been widely treated in recent years; see for example, Särndal (1992), Deville and Särndal (1994), Rao and Shao (1992), Rao (1996) and Shao and Steel (1999). In practice, it is often of interest to estimate the ratio of two population totals, $R = Y/X$, where $(Y, X) = \sum_{i \in U} (y_i, x_i)$, y and x denote two variables of interest potentially missing and U denotes the finite population (of size N) under study. Although variance estimation for a ratio in the presence of imputed data is a problem frequently encountered in practice (especially in business surveys), it has not been, to our knowledge, fully studied in the literature. In this paper, we consider the case Marginal Random Hot Deck (MRHD) imputation performed within the same set of imputation classes for both variables y and x . In other words, to compensate for nonresponse, random hot deck imputation is performed separately for both variables within the same set of imputation classes. This situation occurs frequently in practice. For simplicity, we consider the case of a single imputation class. Extensions to multiple imputation classes are relatively straightforward for most derivations presented in this paper.

In this paper, we derive variance estimators that take sampling, nonresponse and imputation into account. Two distinct frameworks for variance estimation have been studied in the literature: (i) the customary two-phase framework (e.g., Särndal (1992)) and (ii) the reverse framework (e.g., Shao and Steel (1999)). In the two-phase framework, nonresponse is viewed as a second phase of selection. That is, a random sample is selected from the population according to a given sampling design. Then, given the selected sample, the set of respondents is

generated according to the nonresponse mechanism. In the reverse framework, the order of sampling and response is reversed. That is, the population is first randomly divided into a population of respondents and a population of nonrespondents according to the nonresponse mechanism. Then, a random sample is selected from the population (containing respondents and nonrespondents) according to the sampling design. As we will see in section 4, the reverse framework facilitates the derivation of variance estimators but unlike the two-phase framework, it requires the additional assumption that the nonresponse mechanism does not depend on which sample is selected. This assumption is satisfied in many situations encountered in practice. For each framework, inference can be based either on an Imputation Model (IM) or a Nonresponse Model (NM). The IM approach requires the validity of an imputation model, whereas the NM approach requires the validity of a nonresponse model.

In section 2, we introduce notation, assumptions and the imputed estimator of a ratio under weighted MRHD imputation. The IM and NM approaches are then presented in sections 2.1 and 2.2. In section 2.3, the bias of the imputed estimator is discussed. In section 3, variance estimators are derived under the two-phase framework and the IM approach using the method proposed by Särndal (1992). We show that, under MRHD imputation, the naïve variance estimator (that treats the imputed values as observed values) generally overestimates the sampling variance when y and x are positively correlated. In section 4, we derive variance estimators under the reverse framework and both the IM and the NM approaches using the method proposed by Shao and Steel (1999). Finally, we conclude in section 5.

2. Notation and assumptions

Our goal is to estimate R . We select a random sample, s , of size n , according to a given sampling design $p(s)$. A complete-data estimator is given by

$$\hat{R} = \frac{\hat{Y}_{\text{HT}}}{\hat{X}_{\text{HT}}}, \quad (2.1)$$

where $(\hat{Y}_{\text{HT}}, \hat{X}_{\text{HT}}) = \sum_{i \in s} w_i (y_i, x_i)$ denote the Horvitz-Thompson estimators for Y and X , respectively and $w_i = 1/\pi_i$ denotes the sampling weight of unit i , where π_i is its probability of inclusion in the sample. The estimator \hat{R} in (2.1) is asymptotically p -unbiased for R , i.e., $E_p(\hat{R}) \approx R$, where the subscript p denotes the expectation and variance with respect to the sampling design $p(s)$. Since \hat{R} is a nonlinear function of estimated totals, its exact design variance, $V_p(\hat{R})$, cannot be easily obtained. To overcome this problem, Taylor linearization is often applied in order to approximate the exact variance. An asymptotically p -unbiased estimator of the approximate variance of \hat{R} is given by

$$\hat{V}_{\text{SAM}} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} e_i e_j, \quad (2.2)$$

where $e_i = 1/\hat{X}_{\text{HT}} (y_i - \hat{R}x_i)$, $\Delta_{ij} = (\pi_{ij} - \pi_i \pi_j) / \pi_{ij} \pi_i \pi_j$ and π_{ij} is the joint selection probability of units i and j . Note that $\pi_{ii} = \pi_i$. In the case of simple random sampling without replacement, the variance estimator (2.2) reduces to

$$\hat{V}_{\text{SAM}} = \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}^2} [s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}], \quad (2.3)$$

where

$$s_y^2 = \frac{1}{n-1} \sum_{i \in s} (y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum_{i \in s} (x_i - \bar{x})^2$$

and

$$s_{xy} = \frac{1}{n-1} \sum_{i \in s} (x_i - \bar{x})(y_i - \bar{y})$$

with

$$(\bar{y}, \bar{x}) = \frac{1}{n} \sum_{i \in s} (y_i, x_i).$$

We now turn to the case for which both variables x and y may be missing. Let a_i be the response indicator of unit i such that $a_i = 1$ if unit i responds to variable y and $a_i = 0$, otherwise. Similarly, let b_i be the response indicator of unit i such that $b_i = 1$ if unit i responds to variable x and $b_i = 0$, otherwise. Let $s_r^{(y)}$ be the set of respondents to variable y of size r_y and $s_r^{(x)}$ be the set of respondents to variable x of size r_x . Also, let r_{xy} be the number of respondents to both variables y and x . Finally,

let y_i^* and x_i^* denote the imputed values to replace the missing values y_i and x_i , respectively. An imputed estimator of R is given by

$$\hat{R}_I = \frac{\sum_{i \in s} w_i \tilde{y}_i}{\sum_{i \in s} w_i \tilde{x}_i}, \quad (2.4)$$

where $\tilde{y}_i = a_i y_i + (1 - a_i) y_i^*$ and $\tilde{x}_i = b_i x_i + (1 - b_i) x_i^*$. Under weighted MRHD imputation, to compensate for the missing value y_i^* , a donor j is selected at random with replacement from $s_r^{(y)}$ so that

$$P(y_i^* = y_j) = \frac{w_j}{\sum_{i \in s} w_i a_i}.$$

Similarly, to compensate for the missing value x_i^* , a donor k is selected at random with replacement from $s_r^{(x)}$ so that

$$P(x_i^* = x_k) = \frac{w_k}{\sum_{i \in s} w_i b_i}.$$

Note that, when both y_i and x_i are missing, j is generally not equal to k under weighted MRHD imputation.

Random hot-deck imputation within classes is widely used in practice because (i) it preserves the variability of the original data; and (ii) it leads to plausible values. The latter is particularly important in the case of categorical variables of interest. However, random hot-deck imputation within classes suffers from an additional component of variance due to the use of a random imputation mechanism. The main reason weighted MRHD imputation is used is that it leads to asymptotically unbiased estimator under the nonresponse model approach (see section 2.1) unlike unweighted MRHD imputation.

Let $E_I(\cdot | s, s_r^{(y)}, s_r^{(x)})$, $V_I(\cdot | s, s_r^{(y)}, s_r^{(x)})$ and $\text{Cov}_I(\cdot, \cdot | s, s_r^{(y)}, s_r^{(x)})$ denote the conditional expectation, the conditional variance and the conditional covariance operators with respect to the random imputation mechanism (here, weighted MRHD imputation). Using a first-order Taylor expansion, it can be shown that

$$E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) \approx \frac{\bar{y}_r}{\bar{x}_r} \equiv \hat{R}_r, \quad (2.5)$$

where

$$\bar{y}_r = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i a_i}$$

and

$$\bar{x}_r = \frac{\sum_{i \in s} w_i b_i x_i}{\sum_{i \in s} w_i b_i}$$

denote the weighted means of the respondents to variables y and x , respectively. The approximation in (2.5) will be valid if the sample size within classes is sufficiently large, which we assume to be the case. Now, let

$$s_{yr}^2 = \sum_{i \in s} \frac{1}{w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_r)^2$$

and

$$s_{xr}^2 = \sum_{i \in s} \frac{1}{w_i b_i} \sum_{i \in s} w_i b_i (x_i - \bar{x}_r)^2$$

denote the variability of the y -values and the x -values in the set of respondents $s_r^{(y)}$ and $s_r^{(x)}$, respectively. Noting that, under weighted MRHD imputation,

$$V_I(y_i^*) = s_{yr}^2, \quad V_I(x_i^*) = s_{xr}^2$$

and

$$\text{Cov}_I(y_i^*, x_i^*) = 0,$$

we can approximate $V_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)})$ by

$$V_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) \approx \frac{1}{\bar{x}_r^2} \left[\sum_{i \in s} w_i^2 (1 - a_i) s_{yr}^2 + \hat{R}_I^2 \sum_{i \in s} w_i^2 (1 - b_i) s_{xr}^2 \right]. \quad (2.6)$$

Expressions (2.5) and (2.6) will be useful in subsequent sections when discussing the bias and the variance of the imputed estimator \hat{R}_I . As we will see in sections 3 and 4, the conditional variance (2.6) is a measure of the variability due to the imputation mechanism.

Next, we describe two approaches to inference that will be used to obtain variance estimators in sections 3 and 4: the Nonresponse Model (NM) approach and the Imputation Model (IM) approach.

2.1 The nonresponse model approach

In the NM approach, inference is made with respect to the joint distribution induced by the sampling design and the nonresponse model. The nonresponse model is a set of assumptions about the unknown distribution of the response indicators $\mathbf{R}_s = \{(a_i, b_i); i \in s\}$. This unknown distribution is often called the nonresponse mechanism. Let $p_{yi} = P(a_i = 1 | s, \mathbf{Z}_s)$ be the response probability of unit i to variable y , where $\mathbf{Z}_s = \{\mathbf{z}_i; i \in s\}$ and \mathbf{z}_i is a vector of auxiliary variables available for all sample units used to form the imputation classes. Similarly, let $p_{xi} = P(b_i = 1 | s, \mathbf{Z}_s)$ be the response probability of unit i to variable x . We assume that units respond independently; i.e., $p_{yij} = P(a_i = 1, a_j = 1 | s, \mathbf{Z}_s) = p_{yi} p_{yj}$ for $i \neq j$ and $p_{xij} =$

$P(b_i = 1, b_j = 1 | s, \mathbf{Z}_s) = p_{xi} p_{xj}$ for $i \neq j$. However, we do not assume that, for a given unit i , response to variable y is independent of response to variable x . In other words, if we let $p_{xyi} = P(a_i = 1, b_i = 1 | s, \mathbf{Z}_s)$, then we have $p_{xyi} \neq p_{xi} p_{yi}$, in general. Within an imputation class, we assume a uniform response mechanism such that $p_{yi} = p_y$, $p_{xi} = p_x$ and $p_{xyi} = p_{xy}$.

We also assume that, after conditioning on s and \mathbf{Z}_s , the nonresponse mechanism is independent of all other variables involved in the imputed estimator (2.4) as well as the joint selection probabilities. In other words, the distribution of \mathbf{R}_s does not depend on $\mathbf{Y}_s = \{y_i; i \in s\}$, $\mathbf{W}_s = \{w_i; i \in s\}$ and $\mathbf{\Pi}_s = \{\pi_{ij}; i \in s, j \in s\}$, after conditioning on s and \mathbf{Z}_s . As a result, except for the response indicators a_i and b_i , we assume that all the variables involved in the imputed estimator (2.4) as well as the joint selection probabilities are treated as fixed when taking expectations and variances with respect to the nonresponse model. From this point on, we use the subscript q to denote the expectation and variance with respect to the nonresponse mechanism.

2.2 The imputation model approach

In the IM approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design and the nonresponse model. The imputation model is a set of assumptions about the unknown distribution of $(\mathbf{Y}_U, \mathbf{X}_U) = \{(y_i, x_i); i \in U\}$. Within an imputation class, the imputation model, m , in the case of MRHD imputation, is given by

$$m: \begin{cases} y_i = \mu_y + \varepsilon_i \\ x_i = \mu_x + \eta_i \end{cases} \quad (2.7)$$

where ε_i is a random error term such that $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$, for $i \neq j$, $V_m(\varepsilon_i) = \sigma_\varepsilon^2$ and η_i is a random error term such that $E_m(\eta_i) = 0$, $E_m(\eta_i \eta_j) = 0$, for $i \neq j$, $V_m(\eta_i) = \sigma_\eta^2$. Furthermore, we assume that $E_m(\varepsilon_i \eta_i) = \sigma_{\varepsilon\eta}$. Here, $E_m(\cdot)$, $V_m(\cdot)$ and $\text{Cov}_m(\cdot)$ denote respectively the expectation, the variance and the covariance operators with respect to model m . It is implicit in the notation that expectations or variances with respect to model m are conditional on $\mathbf{Z}_U = \{\mathbf{z}_i; i \in U\}$. In this approach, we assume that the distribution of the model errors $(\varepsilon_U, \eta_U) = \{(\varepsilon_i, \eta_i); i \in U\}$ does not depend on s , $s_r^{(y)}$, $s_r^{(x)}$, $\mathbf{W}_U = \{w_i; i \in U\}$ and $\mathbf{\Pi}_U = \{\pi_{ij}; i \in U, j \in U\}$, after conditioning on \mathbf{Z}_U . As a result, except for the variables of interest y and x , all variables involved in the imputed estimator (2.4) are treated as fixed when taking expectations and variances with respect to the imputation model.

2.3 Bias of the imputed estimator

To study the bias of the imputed estimator (2.4), we use the standard decomposition of the total error of \hat{R}_I :

$$\begin{aligned} \hat{R}_I - R &= [\hat{R} - R] + [E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R}] \\ &\quad + [\hat{R}_I - E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)})]. \end{aligned} \quad (2.8)$$

The first term $\hat{R} - R$ on the right-hand side of (2.8) is called the sampling error, the second term $E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R}$ is called the nonresponse error, whereas the third term $\hat{R}_I - E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)})$ is called the imputation error.

Using a first-order Taylor expansion, it can easily be shown that, under the NM approach, the imputed estimator (2.4) is asymptotically pqI -unbiased; that is, $E_{pqI}(\hat{R}_I - R) \approx 0$. Also, under the IM approach and model (2.7), it can be shown that the imputed estimator (2.4) is asymptotically $mpqI$ -unbiased under the IM approach; that is, $E_{mpqI}(\hat{R}_I - R) \approx 0$. Thus, the imputed estimator is robust in the sense that it is valid under either the NM approach or the IM approach. Note that for the asymptotic bias to be equal to 0 under both approaches, we require that the sample size within each imputation class is sufficiently large. From this point on, we thus assume that the bias of \hat{R}_I is negligible.

3. Variance estimation: The two-phase framework

In this section, we derive variance estimators under the two-phase framework and the IM approach according to the method proposed by Särndal (1992) and Deville and Särndal (1994). Using the decomposition (2.8), the total variance of \hat{R}_I can be approximated by

$$\begin{aligned} V_{mpqI}(\hat{R}_I - R) &\approx E_{mpqI}(\hat{R}_I - R)^2 \\ &= V_{\text{SAM}} + V_{\text{NR}} + \tilde{V}_I + 2V_{\text{MIX}}, \end{aligned} \quad (3.1)$$

where $V_{\text{SAM}} = E_m V_p(\hat{R}) = E_m(V_{\text{SAM}})$ is the sampling variance of the complete-data estimator \hat{R} , $V_{\text{NR}} = E_{pq} V_m(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$ is the nonresponse variance of the imputed estimator \hat{R}_I , $\tilde{V}_I = E_{mpqI} V_I(\hat{R}_I - E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) | s, s_r^{(y)}, s_r^{(x)})$ is the imputation variance of the imputed estimator \hat{R}_I , and $V_{\text{MIX}} = E_{pqm}[(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) | s, s_r^{(y)}, s_r^{(x)}]$ is a mixed component. Note that the expression (3.1) contains only one cross product term, $2V_{\text{MIX}}$, because the other cross product terms are all asymptotically equal to 0.

3.1 Estimation of the sampling variance V_{SAM}

Let \hat{V}_{ORD} be the naive variance estimator of \hat{R}_I ; i.e., the variance estimator obtained by treating the imputed values as observed values. The variance estimator \hat{V}_{ORD} is thus obtained by replacing e_i by $\tilde{e}_i = 1/\hat{X}_I(\tilde{y}_i - \hat{R}_I \tilde{x}_i)$ in (2.2) which leads to

$$\hat{V}_{\text{ORD}} = \sum_{i \in S} \sum_{j \in S} \Delta_{ij} \tilde{e}_i \tilde{e}_j. \quad (3.2)$$

As we show now in the case of simple random sampling without replacement, \hat{V}_{ORD} overestimates V_{SAM} under MRHD imputation whenever $\sigma_{\varepsilon\eta} > 0$ (as it is usually the case in practice). After some algebra, we obtain

$$\begin{aligned} E_{ml}(\hat{V}_{\text{ORD}} - \hat{V}_{\text{SAM}} | s, s_r^{(y)}, s_r^{(x)}) \\ \approx \frac{2}{\mu_x^2} \left(\frac{\mu_y}{\mu_x} \right) \left(1 - \frac{n}{N} \right) \left(1 - \frac{r_{xy}}{n} \right) \frac{\sigma_{\varepsilon\eta}}{n}. \end{aligned} \quad (3.3)$$

Expression (3.3) shows that \hat{V}_{ORD} is $mpqI$ -biased for \hat{V}_{SAM} unless $\sigma_{\varepsilon\eta} = 0$, $r_{xy} = n$ (which is the case of complete data) or $n = N$ (which is the census case). The fact that \hat{V}_{ORD} is not a valid estimator of V_{SAM} can be easily explained by noting that although MRHD imputation preserves the variability, s_x^2 and s_y^2 , corresponding to variables x and y , it does not preserve the covariance, s_{xy} , in (2.3). Indeed, imputation tends to underestimate relationships between variables that are positively correlated. As a result, \hat{V}_{ORD} overestimates V_{SAM} because of the presence of the minus sign in front of s_{xy} in (2.3). To overcome this difficulty, Särndal (1992) proposed to estimate $V_{\text{DIF}} = E_{ml}(\hat{V}_{\text{SAM}} - \hat{V}_{\text{ORD}} | s, s_r^{(y)}, s_r^{(x)})$ by a mI -unbiased estimator \hat{V}_{DIF} ; i.e., $E_{ml}(\hat{V}_{\text{DIF}} | s, s_r^{(y)}, s_r^{(x)}) = V_{\text{DIF}}$. However, the derivation of this component for an arbitrary design involves very tedious algebra in the case of a ratio. Therefore, we propose an alternative that does not require any derivation but involves the construction of a new set of imputed values. It can be described as follows: whenever $a_i = 0$ and/or $b_i = 0$, select a donor j at random with replacement from the set of respondents to both variables y and x (i.e., the set of sampled units for which $a_i = 1$ and $b_i = 1$) with probability $w_j / \sum_{i \in S} w_i a_i b_i$ and impute the vector (x_j, y_j) . In other words, whenever one variable is missing, the observed value is discarded and set to missing; the missing values are then replaced by the values of a donor selected at random among the set of respondents to both variables x and y (often called the set of common donors). Similarly, when both variables are missing, the vector (x_j, y_j) of a donor j is imputed. Then, use the standard variance estimator (2.2) valid in the complete response case using these imputed values. Let

\hat{V}_{ORD}^* denote the resulting variance estimator. Note that this new set of imputed values is used only to obtain a valid estimator of the sampling variance and is not used to estimate the parameter of interest R . It can be shown that \hat{V}_{ORD}^* is an asymptotically *mpqI*-unbiased estimator of V_{SAM} . In practice, one could, for example, create a variance estimation file containing the new set of imputed values and use standard variance estimation systems (used in the complete data case) to obtain an estimate of the sampling variance.

3.2 Estimation of the nonresponse variance V_{NR}

An estimator \hat{V}_{NR} of $V_{\text{NR}} = E_{pq} V_m(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$ can simply be obtained by estimating $V_m(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$. Using a first-order Taylor expansion, we obtain

$$\begin{aligned} V_{\text{NR}} \approx & \frac{1}{\mu_x^2} \left\{ \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - 2 \frac{\sum_{i \in s} w_i^2 a_i}{\hat{N} \hat{N}_a} \right] \sigma_e^2 \right. \\ & + \left[\frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - 2 \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N} \hat{N}_b} \right] \left(\frac{\mu_y}{\mu_x} \right)^2 \sigma_\eta^2 \\ & \left. - 2 \left[\frac{\sum_{i \in s} w_i^2 a_i b_i}{\hat{N}_a \hat{N}_b} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i \in s} w_i^2 a_i}{\hat{N} \hat{N}_a} - \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N} \hat{N}_b} \right] \left(\frac{\mu_y}{\mu_x} \right) \sigma_{\varepsilon\eta} \right\}, \quad (3.4) \end{aligned}$$

where $(\hat{N}, \hat{N}_a, \hat{N}_b) = \sum_{i \in s} w_i (1, a_i, b_i)$. Now, let $s_{Ix}^2 = 1/\hat{N} \sum_{i \in s} w_i (\bar{x}_i - \bar{x}_I)^2$ and $s_{Iy}^2 = 1/\hat{N} \sum_{i \in s} w_i (\bar{y}_i - \bar{y}_I)^2$ with $(\bar{x}_I, \bar{y}_I) = 1/\hat{N} \sum_{i \in s} w_i (\bar{x}_i, \bar{y}_i)$. Note that s_{Ix}^2 and s_{Iy}^2 denote respectively the sample variability of the x -values and the y -values after imputation. It can be shown that s_{Ix}^2 and s_{Iy}^2 are respectively asymptotically *ml*-unbiased for the model variances σ_η^2 and σ_e^2 . Also, let $s_{xyr} = 1/\hat{N}_{ab} \sum_{i \in s} w_i a_i b_i (x_i - \bar{x}_{rr})(y_i - \bar{y}_{rr})$, where $\hat{N}_{ab} = \sum_{i \in s} w_i a_i b_i$ and $(\bar{x}_{rr}, \bar{y}_{rr}) = \hat{N}_{ab}^{-1} \sum_{i \in s} w_i a_i b_i (x_i, y_i)$. Note that s_{xyr} is *m*-unbiased for the model covariance $\sigma_{\varepsilon\eta}$. It follows that \hat{V}_{NR} is obtained by estimating the unknown quantities in (3.4), which leads to

$$\hat{V}_{\text{NR}} =$$

$$\begin{aligned} & \frac{1}{\bar{x}_I^2} \left\{ \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}_a^2} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - 2 \frac{\sum_{i \in s} w_i^2 a_i}{\hat{N} \hat{N}_a} \right] s_{Iy}^2 \right. \\ & + \left[\frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}_b^2} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - 2 \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N} \hat{N}_b} \right] \hat{R}_I^2 s_{Ix}^2 \\ & \left. - 2 \left[\frac{\sum_{i \in s} w_i^2 a_i b_i}{\hat{N}_a \hat{N}_b} + \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} - \frac{\sum_{i \in s} w_i^2 a_i}{\hat{N} \hat{N}_a} - \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N} \hat{N}_b} \right] \hat{R}_I s_{xyr} \right\}. \quad (3.5) \end{aligned}$$

The estimator (3.5) is asymptotically *mpqI*-unbiased for V_{NR} . In the special case of simple random sampling without replacement, expression (3.5) reduces to

$$\begin{aligned} \hat{V}_{\text{NR}} = & \frac{1}{\bar{x}_I^2} \left\{ \left(1 - \frac{r_y}{n} \right) \frac{s_{Iy}^2}{r_y} + \hat{R}_I^2 \left(1 - \frac{r_x}{n} \right) \frac{s_{Ix}^2}{r_x} \right. \\ & \left. - 2 \hat{R}_I \left(1 - \frac{r_x r_y}{nr_{xy}} \right) \left(\frac{r_{xy}}{r_x} \right) \left(\frac{r_{xy}}{r_y} \right) \frac{s_{xyr}}{r_{xy}} \right\}. \end{aligned}$$

3.3 Estimation of the imputation variance \tilde{V}_I

An estimator \hat{V}_I of $\tilde{V}_I = E_{mpq} V_I(\hat{R}_I - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$ can simply be obtained by estimating $V_I(\hat{R}_I - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$ given by (2.6). An asymptotically *I*-unbiased of $V_I(\hat{R}_I - \hat{R} | s, s_r^{(y)}, s_r^{(x)})$ is then given by

$$\hat{V}_I = \frac{1}{\bar{x}_I^2} \left[\sum_{i \in s} w_i^2 (1 - a_i) s_{Iy}^2 + \hat{R}_I^2 \sum_{i \in s} w_i^2 (1 - b_i) s_{Ix}^2 \right]. \quad (3.6)$$

It follows that \hat{V}_I in (3.6) is asymptotically *mpqI*-unbiased for \tilde{V}_I . In the special case of simple random sampling without replacement, expression (3.6) reduces to

$$\hat{V}_I = \frac{N^2}{\bar{x}_I^2} \left[\left(1 - \frac{r_y}{n} \right) \frac{s_{Iy}^2}{n} + \hat{R}_I^2 \left(1 - \frac{r_x}{n} \right) \frac{s_{Ix}^2}{n} \right].$$

3.4 Estimation of the mixed component V_{MIX}

Finally, we obtain an estimator \hat{V}_{MIX} of V_{MIX} by estimating

$$E_m[(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) | s, s_r^{(y)}, s_r^{(x)}].$$

Using a first-order Taylor expansion, we obtain

$$\begin{aligned} E_m[(E_I(\hat{R}_I | s, s_r^{(y)}, s_r^{(x)}) - \hat{R})(\hat{R} - R) | s, s_r^{(y)}, s_r^{(x)}] \\ \approx \frac{1}{\mu_x^2} \left\{ \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] \sigma_\varepsilon^2 \right. \\ \left. + \left[\frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] \left(\frac{\mu_y}{\mu_x} \right)^2 \sigma_\eta^2 \right. \\ \left. - 2 \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - 2 \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] \left(\frac{\mu_y}{\mu_x} \right) \sigma_{\varepsilon\eta} \right\}. \quad (3.7) \end{aligned}$$

An estimator of (3.7) is thus given by

$$\begin{aligned} \hat{V}_{\text{MIX}} = \\ \frac{1}{\bar{x}_I^2} \left\{ \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}\hat{N}_a} - \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] s_{Iy}^2 + \left[\frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] \hat{R}_I^2 s_{Ix}^2 \right. \\ \left. - 2 \left[\frac{\sum_{i \in s} w_i^2 a_i}{\hat{N}\hat{N}_a} + \frac{\sum_{i \in s} w_i^2 b_i}{\hat{N}\hat{N}_b} - 2 \frac{\sum_{i \in s} w_i^2}{\hat{N}^2} \right] \hat{R}_I s_{xy} \right\}. \quad (3.8) \end{aligned}$$

The estimator (3.8) is asymptotically *mpqI*-unbiased for V_{MIX} . In the case of simple random sampling without replacement, the component \hat{V}_{MIX} is equal to zero. More generally, the component \hat{V}_{MIX} is equal to zero for any unistage self-weighting design (*i.e.*, a sampling design for which the sampling weight are all equal). For unequal probability designs, it is important to include the component \hat{V}_{MIX} because its contribution (positive or negative) to the overall variance could be substantial (Brick, Kalton and Kim (2004)).

Finally, an asymptotically *mpqI*-unbiased estimator of the total variance $V_{\text{TOT}} = V_{\text{mpqI}}(\hat{R}_I - R)$ is thus given by

$$\hat{V}_{\text{TOT}}^{(\text{TP})} = \hat{V}_{\text{ORD}}^* + \hat{V}_{\text{NR}} + \hat{V}_I + 2\hat{V}_{\text{MIX}}.$$

4. Variance estimation: The reverse framework

In this section, we derive variance estimators under the reverse framework and both the NM and the IM approaches according to the method proposed by Shao and Steel (1999). Recall that, under this framework, we require the additional

assumption that the response probabilities do not depend on the sample s . Under the NM approach, the total variance of \hat{R}_I can be approximated by

$$\begin{aligned} V(\hat{R}_I - R) \approx E_q V_p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) \\ + E_{pq} V_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) + V_q E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b}), \quad (4.1) \end{aligned}$$

where $\mathbf{a} = (a_1, \dots, a_N)'$ and $\mathbf{b} = (b_1, \dots, b_N)'$ denote the vectors of response indicators to variables y and x , respectively.

Under the IM approach, the total variance of \hat{R}_I can be approximated by

$$\begin{aligned} V(\hat{R}_I - R) \approx E_{mq} V_p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) \\ + E_{mpq} V_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) \\ + E_q V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b}). \quad (4.2) \end{aligned}$$

Under both the NM and the IM approaches, an estimator of the first term on the right hand side of (4.1) and (4.2) can be obtained by finding an asymptotically *pl*-unbiased estimator of $V_p E_I(\hat{R}_I | \mathbf{a}, \mathbf{b})$. Also, the second term on the right hand side of (4.1) and (4.2) can be estimated by \hat{V}_I given by (3.6). Under the NM approach, an estimator the last term on the right hand side of (4.1) can be obtained by estimating $V_q E_{pl}(\hat{R}_I | \mathbf{a}, \mathbf{b})$, whereas an estimator of the last term on the right hand side of (4.2) can be obtained by estimating $V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ under the IM approach. As a result, the estimators of the first two terms in (4.1) and (4.2) are identical and thus are valid regardless of the approach (NM or IM) used for inference. Only the third term on the right hand side of (4.1) and (4.2) will depend on the approach used. In the case of the IM approach, specification and validation of the imputation model is crucial to achieve asymptotic unbiasedness of the third component, whereas in the case of the NM approach, the asymptotic unbiasedness of the third component relies of the correct specification of the nonresponse model.

4.1 Estimation of $V_p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$

Using a first-order Taylor expansion and expression (2.5), an estimator of $V_p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$, denoted by \hat{V}_I , is given by

$$\hat{V}_I = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \xi_i \xi_j, \quad (4.3)$$

where

$$\xi_i = \frac{1}{\bar{x}_r} \left[\frac{1}{\hat{N}_a} a_i (y_i - \bar{y}_r) - \left(\frac{\bar{y}_r}{\bar{x}_r} \right) \frac{1}{\hat{N}_b} b_i (x_i - \bar{x}_r) \right].$$

In other words, the estimator \hat{V}_1 is obtained from the complete data variance estimator (2.2) by replacing e_i by ξ_i . In the case of simple random sampling without replacement, the estimator (4.3) reduces to

$$\hat{V}_1 = \left(1 - \frac{n}{N}\right) \frac{1}{\bar{x}_r^2} \left[\frac{s_{yr}^2}{r_y} + \hat{R}_r^2 \frac{s_{xr}^2}{r_x} - 2\hat{R}_r \left(\frac{r_{xy}}{r_x} \right) \left(\frac{r_{xy}}{r_y} \right) \frac{s_{xyr}}{r_{xy}} \right]. \quad (4.4)$$

4.2 Estimation of $V_q E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ under the NM approach

First, note that

$$E_{pl}(\hat{R}_I) \approx \frac{Y_a N_b}{N_a X_b},$$

where $(Y_a, N_a) = \sum_{i \in U} a_i(y_i, 1)$ and $(X_b, N_b) = \sum_{i \in U} b_i(x_i, 1)$.

Using a first-order Taylor expansion, it can be shown that $V_q E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ can be approximated by

$$V_q E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) \approx \frac{1}{N\bar{X}^2} \left[\left(\frac{1 - p_y}{p_y} \right) S_y^2 + R^2 \left(\frac{1 - p_x}{p_x} \right) S_x^2 - 2R \left(\frac{p_{xy} - p_x p_y}{p_x p_y} \right) S_{xy} \right], \quad (4.5)$$

where

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1} \sum_{i \in U} (x_i - \bar{X})^2$$

and

$$S_{xy} = \frac{1}{N-1} \sum_{i \in U} (x_i - \bar{X})(y_i - \bar{Y})$$

with

$$(\bar{Y}, \bar{X}) = \frac{1}{N} \sum_{i \in U} (y_i, x_i).$$

An estimator of $V_q E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ is obtained by estimating unknown quantities in (4.5), which leads to

$$\begin{aligned} \hat{V}_2^{(NM)} &= \frac{1}{\hat{N}\bar{x}_I^2} \left[\left(\frac{1 - \hat{p}_y}{\hat{p}_y} \right) s_{by}^2 + \hat{R}_I^2 \left(\frac{1 - \hat{p}_x}{\hat{p}_x} \right) s_{bx}^2 - 2\hat{R}_I \left(\frac{\hat{p}_{xy} - \hat{p}_x \hat{p}_y}{\hat{p}_x \hat{p}_y} \right) s_{xyr} \right] \\ &= \frac{1}{\bar{x}_I^2} \left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}} \right) s_{by}^2 + \hat{R}_I^2 \left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}} \right) s_{bx}^2 - 2\hat{R}_I \left(\frac{\hat{N}_{ab}}{\hat{N}_b \hat{N}_b} - \frac{1}{\hat{N}} \right) s_{xyr} \right], \quad (4.6) \end{aligned}$$

$$\text{where } \hat{p}_y = \frac{\hat{N}_a}{\hat{N}}, \hat{p}_x = \frac{\hat{N}_b}{\hat{N}} \text{ and } \hat{p}_{xy} = \frac{\hat{N}_{ab}}{\hat{N}}.$$

The estimator (4.6) is asymptotically pqI -unbiased for the approximate variance (4.5), noting that s_{by}^2 , s_{bx}^2 and s_{xyr} are asymptotically pqI -unbiased for S_y^2 , S_x^2 and S_{xy} , respectively. In the case of simple random sampling without replacement, the estimator (4.6) reduces to

$$\begin{aligned} \hat{V}_2^{(NM)} &= \frac{1}{\bar{x}_I^2} \left[\left(\frac{n}{N} - \frac{r_y}{N} \right) \frac{s_{by}^2}{r_y} + \hat{R}_I^2 \left(\frac{n}{N} - \frac{r_x}{N} \right) \frac{s_{bx}^2}{r_x} - 2\hat{R}_I \left(\frac{n}{N} - \frac{r_{xy}}{N} \right) \left(\frac{r_{xy}}{r_x} \right) \left(\frac{r_{xy}}{r_y} \right) \frac{s_{xyr}}{r_{xy}} \right]. \quad (4.7) \end{aligned}$$

4.3 Estimation of $E_q V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ under the IM approach

Using a first-order Taylor expansion, it can be shown that $E_q V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ can be approximated by

$$\begin{aligned} E_q V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) &\approx \frac{1}{\mu_x^2} \left[\left(\frac{1}{E_q(N_a)} - \frac{1}{N} \right) \sigma_e^2 + \left(\frac{\mu_y}{\mu_x} \right)^2 \left(\frac{1}{E_q(N_b)} - \frac{1}{N} \right) \sigma_n^2 - 2 \left(\frac{\mu_y}{\mu_x} \right) \left(E_q \left(\frac{N_{ab}}{N_b N_b} \right) - \frac{1}{N} \right) \sigma_{en} \right], \quad (4.8) \end{aligned}$$

where $N_{ab} = \sum_{i \in U} a_i b_i$. An estimator of $E_q V_m E_{pl}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ is obtained by estimating unknown quantities in (4.8), which leads to

$$\begin{aligned} \hat{V}_2^{(IM)} &= \frac{1}{\bar{x}_I^2} \left[\left(\frac{1}{\hat{N}_a} - \frac{1}{\hat{N}} \right) s_{by}^2 + \hat{R}_I^2 \left(\frac{1}{\hat{N}_b} - \frac{1}{\hat{N}} \right) s_{bx}^2 - 2\hat{R}_I \left(\frac{\hat{N}_{ab}}{\hat{N}_b \hat{N}_b} - \frac{1}{\hat{N}} \right) s_{xyr} \right]. \quad (4.9) \end{aligned}$$

The estimator (4.9) is asymptotically $mpqI$ -unbiased for the approximate variance (4.8). It is interesting to note that, under weighted MRHD imputation, the estimator $\hat{V}_2^{(NM)}$ in (4.6) obtained under the NM approach is identical to $\hat{V}_2^{(IM)}$ in (4.9) obtained under the IM approach. However, this may not be the case with a different imputation method. Also, the component \hat{V}_2 is negligible with respect to \hat{V}_1 when the sampling fraction n/N is negligible, where \hat{V}_2 stands for $\hat{V}_2^{(NM)}$ or $\hat{V}_2^{(IM)}$. In this case, the component \hat{V}_2 may be omitted from the calculations.

Finally, an estimator of the total variance under the reverse framework is given by

$$\hat{V}_{TOT}^{(RE)} = \hat{V}_1 + \hat{V}_I + \hat{V}_2.$$

Under the reverse framework, both the NM approach and the IM approach lead to the same estimator of the total variance. Thus, the variance estimator $\hat{V}_{\text{TOT}}^{(\text{RE})}$ is robust in the sense that it is valid under either the NM approach or the IM approach.

5. Summary and conclusions

In this paper, we have derived variance estimators for the imputed estimator of a ratio under two different frameworks. The reverse framework facilitates the derivation of the variance expressions (in comparison with the customary two-phase framework), especially if the sampling fraction is small, in which case we can omit the component \hat{V}_2 . However, unlike the two-phase framework, it requires an additional assumption that the response probabilities do not depend on the realized sample s . Also, the two-phase framework uses a natural decomposition of the total error that leads to a natural decomposition of the total variance. That is, the total variance can be expressed as the sum of the sampling variance, the nonresponse variance and the imputation variance, which allows the survey statistician to get an idea of the relative magnitude of each component. Under the reverse approach, there is no easy interpretation for the variance components (except the imputation variance).

We have considered the case of weighted MRHD imputation within classes. Another version of weighted random hot-deck imputation, which we call weighted joint random hot deck (JRHD) imputation, is identical to weighted MRHD imputation, except that when both variables are missing, a donor j is selected at random from the set of common donors (*i.e.*, the set of respondents to both variables y and x) with probability $w_j / \sum_{i \in s} w_i a_i b_i$ and the vector (x_j, y_j) is imputed. This version of the method helps preserving relationships between survey

variables, contrary to imputing independently each variable. The results for JRHD imputation can be obtained using similar techniques presented in this paper. Finally, the results presented in this paper can be easily extended to the case of both deterministic and random regression imputation performed within imputation classes.

Acknowledgements

The author thanks an Associate Editor for constructive comments and suggestions that helped improving the quality of the paper. David Haziza's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

- Brick, M.J., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.
- Deville, J.-C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.
- Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.
- Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.
- Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Efficient bootstrap for business surveys

James Chipperfield and John Preston¹

Abstract

The Australian Bureau of Statistics has recently developed a generalized estimation system for processing its large scale annual and sub-annual business surveys. Designs for these surveys have a large number of strata, use Simple Random Sampling within Strata, have non-negligible sampling fractions, are overlapping in consecutive periods, and are subject to frame changes. A significant challenge was to choose a variance estimation method that would best meet the following requirements: valid for a wide range of estimators (*e.g.*, ratio and generalized regression), requires limited computation time, can be easily adapted to different designs and estimators, and has good theoretical properties measured in terms of bias and variance. This paper describes the Without Replacement Scaled Bootstrap (WOSB) that was implemented at the ABS and shows that it is appreciably more efficient than the Rao and Wu (1988)'s With Replacement Scaled Bootstrap (WSB). The main advantages of the Bootstrap over alternative replicate variance estimators are its efficiency (*i.e.*, accuracy per unit of storage space) and the relative simplicity with which it can be specified in a system. This paper describes the WOSB variance estimator for point-in-time and movement estimates that can be expressed as a function of finite population means. Simulation results obtained as part of the evaluation process show that the WOSB was more efficient than the WSB, especially when the stratum sample sizes are sometimes as small as 5.

Key Words: Variance; Bootstrap; Stratified sampling.

1. Introduction

In 2000, the Australian Bureau of Statistics (ABS) first obtained a register of businesses containing taxation data from the Australian Taxation Office (ATO). The data items included turnover, sales, and other expense items. In 2001, the ABS used this register as a sampling frame for some surveys in order to improve the efficiency of its sample designs. This data is updated for each business at least annually. To make maximum use of these administrative data items in estimation the ABS developed a generalized estimation system called ABSEST, with the capability of supporting generalized regression estimation (GREG) and variance estimation. ABSEST has been routinely used for the monthly ABS Retail Survey since July 2005.

A generalized estimation system is highly desirable for statistical agencies as it supports a variety of survey output requirements at high levels of statistical rigor for an acceptable cost. The ABS has invested considerable resources into its generalized estimation system for business surveys. Prior to 1998, the ABS's generalized estimation system was capable of Horvitz-Thompson, ratio, and two-phase estimation with variance estimates based on Taylor Series (TS) approximations. In 1999, the Taylor Series method was replaced with the Jackknife method. Subsequent feedback about the computer design and usability were that changes to the generalized estimation system made it increasingly complex to maintain and develop and that processing time could be undesirably long. These key features were important when choosing the variance estimation method for ABSEST.

Core survey output statistics for ABS business surveys are estimates at a point in time, estimates of movement between two time points, and estimates of rates. Business surveys are equal probability designs within stratum, are highly stratified (100s of strata), can be either single or two phase sample designs, and for surveys that sample on more than one occasion the overlapping sample can range from 0 to 100%. The sample size for business surveys range from less than 1,000 to 15,000; stratum level sample sizes can be as low as 3 and as high as several hundred.

Section 2 introduces the GREG estimator. Section 3 discusses alternative variance estimators for GREG and justifies why the Bootstrap variance estimator was chosen for ABSEST. Section 4 describes the Without Replacement Scaled Bootstrap (WOSB) and Rao and Wu (1988)'s With Replacement Bootstrap (WSB) variance estimators for point-in-time estimates under single-phase designs. Section 5 describes the WOSB for movement estimates. Section 6 measures the bias and variance properties of WOSB and WSB in a simulation study. Section 7 gives some concluding remarks.

2. Generalised regression (GREG) estimator

In this section we briefly describe the GREG that is implemented in ABSEST. Consider a finite population U divided into H strata $U = \{U_1, U_2, \dots, U_H\}$, where U_h is comprised of N_h units. The finite population total of interest is $Y = \sum_h Y_h$, where $Y_h = \sum_{i \in U_h} y_{hi}$ and $h = 1, \dots, H$. Within stratum h , the sample s_h of n_h units is selected

1. James Chipperfield, Australian Bureau of Statistics. E-mail: james.chipperfield@abs.gov.au; John Preston, Australian Bureau of Statistics. E-mail: john.preston@abs.gov.au.

from U_h by Simple Random Sampling without Replacement (SRSWOR). The complete sample set is denoted by $s = \{s_1, s_2, \dots, s_H\}$.

Consider the case where a K vector of auxiliary variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ki}, \dots, x_{Ki})^T$ is available for $i \in s$ and the corresponding vector of population totals $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$ are known. The GREG estimator (Särndal, Swensson and Wretman 1992, page 227) is given by $\hat{Y}_{\text{reg}} = \sum_{i \in s} \tilde{w}_i y_i = \sum_{i \in s} w_i y_i + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}}$, where $\tilde{w}_i = w_i g_i$, $w_i = N_h / n_h$, $\hat{\mathbf{B}} = \hat{\mathbf{T}}^{-1} \hat{\mathbf{t}}$ with $\hat{\mathbf{T}}^{-1}$ being the generalised inverse of $\hat{\mathbf{T}}$, $\hat{\mathbf{X}} = \sum_{i \in s} w_i \mathbf{x}_i$ and $g_i = (1 + \sigma_i^{-2} \mathbf{x}_i^T \hat{\mathbf{T}}^{-1} (\mathbf{X} - \hat{\mathbf{X}})^T)$, $\hat{\mathbf{T}} = \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i^T \sigma_i^{-2}$, $\hat{\mathbf{t}} = \sum_{i \in s} w_i \mathbf{x}_i y_i \sigma_i^{-2}$, σ_i^2 is a constant motivated by the superpopulation model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ such that ε_i is independently and identically distributed with mean 0 and variance σ_i^2 , and $E(\hat{\mathbf{B}}) = \boldsymbol{\beta}$. It is well known that \hat{Y}_{reg} is unbiased to $O(n^{-1})$. The weights \tilde{w}_i are stored for ready calculation of estimates. In practice bounds will be placed on the weights, \tilde{w}_i . If the weights, \tilde{w}_i , given by the above equation, are outside these bounds, they are calculated through iteration (see Method 5 of Singh and Mohl 1996).

The expression for \hat{Y}_{reg} can be adapted to a range of estimates, including domains and multi-phase (see Esteveao, Hidioglou and Särndal 1995). For example, when $\mathbf{x}_i = 1$, \hat{Y}_{reg} becomes the Horvitz-Thompson estimator given by $\hat{Y} = \sum_h w_h \sum_{i \in s_h} y_{hi}$ with estimated variance $\widehat{\text{Var}}(\hat{Y}) = \sum_h N_h^2 n_h^{-1} (1 - f_h) \hat{s}_h^2$, where $\hat{s}_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_{hi} - \hat{y}_h)^2$, $\hat{y}_h = \sum_{i \in s_h} y_{hi} / n_h$ and $f_h = n_h / N_h$.

3. Comparison of alternative variance estimators

The ABSEST variance estimation method was required to have bias and variance properties that were competitive in simulation studies, when compared with alternatives in the literature. In order to simplify the maintenance and development of the system, the variance estimation system specifications were required to be generic such that all calculations were largely independent of the estimator. (ABSEST need only support SRSWOR within stratum and single stage sample designs). Also, strong consideration was given to minimise the computational costs.

Firstly, we also considered the Bootstrap, Jackknife and Balanced Repeated Replication (BRR) methods (Shao and Tu 1995, Rao and Wu 1988). Consider estimating the variance of a function $\hat{\theta} = \theta(\hat{\mathbf{Y}})$, where $\hat{\mathbf{Y}}$ is a P vector of estimates $\hat{\mathbf{Y}} = \sum_{i \in s} w_i y_i$, y_i is the response vector from unit i with elements y_{pi} , and θ is a smooth function. Estimating the variance using a replication method involves the following steps:

- (i) independently sub-sampling from the set s a total of R times;

- (ii) for each of the R sub-samples computing $w_i^* = b_i^* w_i$, where b_i^* depends upon the number of times unit i is selected in the sub-sample;

- (iii) calculate $\hat{\theta}^* = \hat{\theta}(\hat{\mathbf{Y}}^*)$, where $\hat{\mathbf{Y}}^* = \sum_{i \in s} w_i^* y_i$;

- (iv) estimate variance of θ by $\widehat{\text{Var}}_{\text{rep}}(\hat{\theta}) = (R - 1)^{-1} \sum_{r=1}^R (\hat{\theta}^{*(r)} - \hat{\theta})^2$, where $\hat{\theta}^{*(r)}$ is the estimate of θ based on the r^{th} replicate sample. Note: the expression for replicate weights, $w_i^* = b_i^* w_i$, includes the Jackknife, Bootstrap and Balanced Repeated Replication as special cases.

As we can express \hat{Y}_{reg} by a function $\hat{\theta}$, the variance of \hat{Y}_{reg}^* can be calculated by the above steps where specifically steps (iii) and (iv) respectively become: (iii) calculate $\hat{Y}_{\text{reg}}^* = \sum_{i \in s} \tilde{w}_i^* y_i$, where $\tilde{w}_i^* = w_i^* g_i^*$ and g_i^* has the same form as g_i but is calculated using the weights w_i^* instead of the weights w_i for $i \in s$; (iv) estimating variance by $\widehat{\text{Var}}_{\text{rep}}(\hat{Y}_{\text{reg}}) = (R - 1)^{-1} \sum_{r=1}^R (\hat{Y}_{\text{reg}}^{*(r)} - \hat{Y}_{\text{reg}})^2$.

The attractive feature of these replication methods is that only the selection of the replicate samples and the value b_i^* is required to calculate unbiased variance estimates for many commonly used sample designs and for estimators that have good first order Taylor Series approximations. Also if the replicate weights, \tilde{w}_i^* are stored the variance estimates of \hat{Y}_{reg}^* require simple calculations that can be completed in a short time; this approach of storing replicate weights has been applied successfully by the ABS' generalized estimation system for household surveys. Once replicate weights are available, calculation of variance for a variety of analysis, such as linear regression, involves simple calculations that require little time and does not require the analyst to have knowledge about the sample design. Next we consider the relative merits of some replicate variance estimators for implementation in ABSEST.

The drop-one Jackknife forms replicate samples, s^* , by dropping one unit at a time. This implies that $R = n$. For large-scale surveys this storage requirement is excessive. The delete-a-group Jackknife, while reducing R by dropping a group of units within a stratum at a time, would still have at least $R = 2H$ replicates- a minimum of two groups per stratum is required to calculate variance. Despite performing well in an empirical study where $n_h = 2$ (see Shao and Tu 1995, page 251), the Jackknife was rejected on the basis of its excessive storage requirement.

For stratified designs the scaled Balanced Repeated Replication (BRR) requires approximately $R = H$ replicate weights. Firstly, the replicate samples are formed by randomly splitting the stratum sample s_h into two groups then allocating one of these groups to s_h^* for each

$h = 1, \dots, H$. The allocation of groups to replicates, defined by a Hadamard matrix, is done in such a way to eliminate between stratum covariance in the replicate samples. The Grouped BRR (GBRR) (see Shao and Tu 1995) can arbitrarily reduce R at the cost of introducing between stratum covariance in the replicate samples. Preston and Chipperfield (2002) showed in an empirical evaluation for a typical ABS business survey that BRR (and GBRR) was significantly more unstable than the Bootstrap.

In their summary of the literature, Kovar, Rao and Wu (1988) found that the scaled Bootstrap tended to have a larger bias compared with the Jackknife or TS when estimating the variance of GREG estimates. As the relative assessment of these methods varied according to the underlying simulation model and the stratum sample size it was important to make an assessment that was based on a model and sample design that were typical of ABS business surveys. Section 6 shows these properties to be acceptable. Unlike the other replication methods, the value of R for the Bootstrap may be chosen arbitrarily and so meet storage and computation restrictions. Further, the selection of the Bootstrap replicate samples is more easily specified in a computer system compared with selection of the BRR replicate samples.

We considered the relative merits of a number of other variance estimators for implementation in ABSEST. The TS method was not suitable as its variance expression for complex estimands involves many terms specific to the estimand, making it difficult to adapt into a generalized system. This problem was addressed by Nordberg (2000) who described a method for variance calculation that automatically generates the Taylor Series expansions. They implemented the method in a computer system, called CLAN. CLAN can handle any function of means under Probability Proportional to Size (PPS) and cluster sampling. A limitation of CLAN is that it does not produce replicate weights which support complex analysis, such as regression analysis, either within or outside a statistical agency; it requires knowledge of the sample design; and it would be a relatively complex system to specify and maintain in a generalized system (in comparison to the Bootstrap). For the same reason other linearized variance estimators (described in Estevao, Hidioglou and Särndal 1995 and evaluated in Yung and Rao 1996) were rejected, despite good theoretical properties, good empirical results and being computationally efficient.

On the above considerations, the preferred variance estimation method for ABSEST was the Bootstrap. In the next section we describe the WOSB and WSB, where only the former is implemented in ABSEST.

4. Without replacement scaled bootstrap (WOSB) for point in time estimates

4.1 Method

For point-in-time GREG estimates, the Without Replacement Scaled Bootstrap (WOSB) variance estimator involves repeating the following R times:

- (a) forming the set s^r by selecting m_h units by SRSWOR from s_h independently within each stratum $h = 1, \dots, H$, where $m_h = \lfloor n_h/2 \rfloor$ and the operator $\lfloor \cdot \rfloor$ rounds down its argument down to the nearest integer;
- (b) calculating $w_{hi}^* = w_{hi}(1 - \gamma_h + \gamma_h n_h / m_h \delta_{hi}^*)$ for $i \in s_h$, where $\gamma_h = \sqrt{(1 - f_h)m_h / (n_h - m_h)}$, δ_{hi}^* is 1 if $i \in s_h^r$ and 0 otherwise; and
- (c) calculating $\tilde{w}_{hi}^* = w_{hi}^* g_{hi}^*$ for $i \in s$; and
- (d) calculating the r^{th} Bootstrap estimate of Y , $\hat{Y}_{\text{reg}}^* = \sum_{i \in s} \tilde{w}_{hi}^* y_i$. The justification $m_h = \lfloor n_h/2 \rfloor$ is given in section 4.2. The Bootstrap variance estimator is given by the Monte Carlo approximation, $\text{Var}_B(\hat{Y}_{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^R (\hat{Y}_{\text{reg}}^{*(r)} - \hat{Y}_{\text{reg}})^2$. The WSB method is the same as WOSB except that the replicate samples are selected by SRSWR and the scaling factor is instead $\gamma_h = \sqrt{(1 - f_h)m_h / (n_h - 1)}$, where m_h is often set to $n_h - 1$ in the literature. Preston and Chipperfield (2002) found that WOSB was found to have significantly less replication error than the WSB- the error due to replicate sampling and conditional on the sample set.

It is easy to see that the WOSB and WSB estimators are unbiased estimators of $\text{Var}(\hat{\theta})$. The TS approximate variance is given by $\widehat{\text{Var}}(\hat{\theta}) = \nabla' \hat{\theta} \hat{V}(\hat{Y}) \nabla \hat{\theta}$, where $\hat{V}(\hat{Y})$ is a $P \times P$ matrix with elements

$$\widehat{\text{Cov}}(\hat{Y}_p, \hat{Y}_{p'}) = \frac{N^2(1-f)}{n} \hat{s}_{p,p'}$$

where

$$\hat{s}_{p,p'} = \frac{1}{n-1} \sum_{i \in s} (y_{pi} - \hat{y}_p)(y_{p'i} - \hat{y}_{p'});$$

$$\hat{y}_p = \frac{1}{n} \sum_{i \in s} y_{pi};$$

$$\hat{Y}_p = \sum_{i \in s} w_i y_{pi}$$

for $p, p' = 1, \dots, P$, and $\nabla' = (\partial/\partial Y_1, \dots, \partial/\partial Y_P)|_{\hat{Y}}$. It is easy to see that

$$E_*(\widehat{\text{Var}}(\hat{\theta}^*)) = \nabla \hat{\theta}' E_*(\hat{V}(\hat{Y}^*)) \nabla \hat{\theta} = \nabla' \hat{\theta} \hat{V}(\hat{Y}) \nabla \hat{\theta},$$

by noting that

$$E_*[\widehat{\text{Cov}}(\hat{Y}_p^*, \hat{Y}_{p'}^*)] = \widehat{\text{Cov}}(\hat{Y}_p, \hat{Y}_{p'})$$

where E_* denotes the expectation with respect to re-sampling. Note the scaling constants applied to w_{hi} to calculate the replicate weights are chosen so that the correct finite population correction factor is obtained. It therefore follows that the Monte Carlo approximation to the variance, $\widehat{\text{Var}}_B(\hat{\theta}) = (R-1)^{-1} \sum_{r=1}^R (\hat{\theta}^{*(r)} - \hat{\theta})^2$, is unbiased for $\text{Var}(\hat{\theta})$.

4.2 A note on the relative efficiency of WSB and WOSB sampling

To simplify notation, let $\hat{v}_{\text{boot}} = \widehat{\text{Var}}_B(\hat{\theta})$. The variance of the Bootstrap variance estimator can be written as

$$\text{Var}(\hat{v}_{\text{boot}}) = \text{Var}_s(E_*[\hat{v}_{\text{boot}} | s]) + E_*[\text{Var}_*(\hat{v}_{\text{boot}} | s)],$$

where s denotes the expectation with respect to the sample design. If \hat{v}_{boot} is unbiased (i.e., $E_*[\hat{v}_{\text{boot}} | s] = \text{Var}(\hat{\theta})$) then $\text{Var}(\hat{v}_{\text{boot}})$ does not depend upon how the replicate samples are selected. The term $\text{Var}_*(\hat{v}_{\text{boot}} | s)$ is the replication error conditional on the sample and is inversely proportional to R . The value of R is chosen to be sufficiently large such that $\text{Var}_s(E_*[\hat{v}_{\text{boot}} | s])$ is small relative to \hat{v}_{boot} , the estimated sample variance. The efficiency of two Bootstrap estimators can be compared by the size of $\text{Var}_s(E_*[\hat{v}_{\text{boot}} | s])$ when both estimators have the same value of R . Next we summarise empirical results based on actual data that show the WOSB can be significantly more efficient than WSB. The benefits of efficiency are either reduced computation time and/or more accurate variance estimates.

Preston and Chipperfield (2002) compared the efficiency of WOSB with $m_h = [n_h/2]$ and WSB with $m_h = n_h - 1$ (see Rao and Wu 1984) for the Australian Quarterly Economic Activity Survey in March 2000. This survey has a stratum level sample size that varies from 4 and into the 100s. The results (derived from Preston and Chipperfield 2002, Table 1) show that at the national level the size of $\text{Var}_s(E_*[\hat{v}_{\text{boot}} | s])$ was 54% smaller for WOSB compared with WSB sampling when $R = 100$ (See Preston and Chipperfield 2002 for more empirical estimates of $\text{Var}_s(E_*[\hat{v}_{\text{boot}} | s])$ for WSB and WOSB). In other words, WOSB required about half the number of replicates to achieve the same replication error as WSB. This represents a significant efficiency gain. Another benefit of WOSB over WSB is that the computational time in selecting the replicate samples is considerably less.

From empirical investigations, the choice of $m_h = [n_h/2]$ for WOSB minimized $\text{Var}_s(E_*[\hat{v}_{\text{boot}} | s])$. As n increases, we suspect that the difference between WOSB and WSB will reduce to approximately zero. More work needs to be done to establish these properties.

5. Movement variance between single phase estimates

A key output requirement of many business surveys is the estimate of change between two time points. Denote the finite population at time t by $U^{(t)} = \{U_1^{(t)}, U_2^{(t)}, \dots, U_H^{(t)}\}$, where $U_h^{(t)}$ is the stratum h population at time t that is made up of $N_h^{(t)}$ units. The population total at time t is $Y^{(t)} = \sum_h \sum_{i \in U_h^{(t)}} y_{hi}^{(t)}$. Estimating the variance of $\Delta^{(t)} = \hat{Y}^{(t)} - \hat{Y}^{(t-1)}$, the difference between two time periods, is the focus of this section. The terms corresponding to n_h , f_h and s_h^2 at time t are denoted by $n_h^{(t)}$, $f_h^{(t)}$ and $s_h^{(t)2}$ respectively. When sampling on two occasions define N_c , n_{hc} , $n_{ch}^{(1)}$ and $n_{ch}^{(2)}$ to be the number of units in the following sets $U_h^{(1)} \cap U_h^{(2)}$, $s_h^{(c)} = s_h^{(1)} \cap s_h^{(2)}$, $s_{h\bar{c}}^{(1)} = s_h^{(1)} - s_h^{(c)}$ and $s_{h\bar{c}}^{(2)} = s_h^{(2)} - s_h^{(c)}$ respectively. In ABS business surveys the time 1 sample of size $n_h^{(1)}$ is an SRSWOR from $U_h^{(1)}$. The time 2 sample is the union of the following two samples: an SRSWOR of n_{hc} units from $s_h^{(c)}$ and an SRSWOR of $n_{ch}^{(2)}$ units from $U_h^{(2)} - (U_h^{(1)} \cap U_h^{(2)})$. The time 2 sample is effectively an SRSWOR from $U_h^{(2)}$. At the ABS, the size of the overlapping sample, n_{hc} , is controlled by the Permanent Random Number method (see Brewer, Gross and Lee 1999).

The estimator of $\text{Var}(\hat{\Delta})$ can be expressed as

$$\text{Var}(\hat{\Delta}) = \text{Var}(\hat{Y}^{(1)}) + \text{Var}(\hat{Y}^{(2)}) - 2\text{Cov}(\hat{Y}_1^{(1)}, \hat{Y}_2^{(2)}).$$

Consider the Horvitz-Thompson estimator $\hat{\Delta} = \hat{Y}^{(2)} - \hat{Y}^{(1)}$, where $t = 1, 2$ and \hat{Y}^t is defined analogously to \hat{Y} . Tam (1985) show that when $U_h^{(1)} = U_h^{(2)}$, an unbiased estimator of $\text{Var}(\hat{\Delta})$ under the above sampling scheme is

$$\widehat{\text{Var}}(\hat{\Delta}) = \widehat{\text{Var}}(\hat{Y}^{(1)}) + \widehat{\text{Var}}(\hat{Y}^{(2)}) - 2\widehat{\text{Cov}}(\hat{Y}^{(1)}, \hat{Y}^{(2)}),$$

where

$$\widehat{\text{Var}}(\hat{Y}^{(t)}) = \sum_h N_h^{(t)} (1 - f_t) s_h^{(t)2} / n_h^{(t)},$$

$$\widehat{\text{Cov}}(Y^{(1)}, Y^{(2)}) = \sum_h N_h^{(1)} (1 - f_{12,h}) s_h^{(12)} n_{hc} / (n_h^{(1)} n_h^{(2)}),$$

$$s_h^{(12)} = (n_{hc} - 1)^{-1} \sum_{i \in s_h^{(c)}} (y_{1i} - \hat{y}_1)(y_{2i} - \hat{y}_2),$$

$$\hat{y}_t = n_{hc}^{-1} \sum_{i \in s_h^{(c)}} y_{ti}$$

for $t = 1, 2$ and $f_{12,h} = n_h^{(1)} n_h^{(2)} / n_{hc} N_h$.

When $U_h^{(1)} \neq U_h^{(2)}$, a more general form of Tam's estimator is given by $\widehat{\text{Var}}(\hat{\Delta})$, except that

$$\widehat{\text{Var}}(\hat{Y}^{(t)}) = \sum_h N_h^{(t)2} (1 - f_t) s_h^{(t)2} / n_h^{(t)},$$

$$\widehat{\text{Cov}}(\hat{Y}^{(1)}, \hat{Y}^{(1)}) = \sum_h N_h^{(1)} N_h^{(2)} / (n_h^{(1)} n_h^{(2)}) n_c (1 - f_{12,h}) s_h^{(12)}$$

and

$$f_{12,h} = \frac{n_h^{(1)} n_h^{(2)} N_{hc}}{n_{hc} N_h^{(1)} N_h^{(2)}}.$$

For the reminder of this section we assume that $\widehat{\text{Var}}(\hat{\Delta})$ is unbiased for $\text{Var}(\hat{\Delta})$ when $U_h^{(1)} \neq U_h^{(2)}$. (It is worthwhile noting that $\widehat{\text{Var}}(\hat{\Delta})$ can take negative values when $U_h^{(1)} \neq U_h^{(2)}$. Nordberg (2000) gives an unbiased estimator of $\text{Var}(\hat{\Delta})$ for the regression estimator when $U_h^{(1)} \neq U_h^{(2)}$, but there is no obvious way in which it can be used with the Bootstrap as described in this paper.)

Estimating the variance of $\hat{\Delta}_{\text{reg}} = \hat{Y}_{\text{reg}}^{(1)} - \hat{Y}_{\text{reg}}^{(2)}$, the movement between GREG estimates at times 1 and 2, using WOSB involves repeating the following R times:

- forming the set s^* by independently selecting $m_{ch} = [n_{ch}/2]$, $m_{ch}^{(1)} = [n_{ch}^{(1)}/2]$ and $m_{ch}^{(2)} = [n_{ch}^{(2)}/2]$ units by SRSWOR from the sets s_{hc} , $s_{hc}^{(1)}$, and $s_{hc}^{(2)}$ respectively;
- for $i \in s_h^{(1)}$ calculate the replicate weights

$$w_{hi}^{*(1)} = N/n_h^{(1)} \left[1 - \gamma_{ch} \frac{n_{ch}}{n_h^{(1)}} - \gamma_{1ch} \frac{n_{hc}^{(1)}}{n_h^{(1)}} + \gamma_{ch} \frac{n_{ch}}{m_{ch}} \delta_{hi}^{r(1)} \right]$$

for $i \in s_{hc}$,

$$w_{hi}^{*(1)} = \left[1 - \gamma_{ch} \frac{n_{ch}}{n_{1h}} - \gamma_{1ch} \frac{n_{1ch}}{n_{1h}} + \gamma_{1ch} \frac{n_{1ch}}{m_{1ch}} \delta_{hi}^{r(1)} \right]$$

for $i \in s_{hc}^{(1)}$, where

$$\gamma_{1ch} = \sqrt{\frac{[n_{1h}(1 - f_h) - n_{ch}(1 - f_{12,h})] m_{1ch}}{\{n_{1ch}(n_{1ch} - m_{1ch})\}}},$$

$$\gamma_{ch} = \sqrt{(1 - f_{12,h}) m_{ch} / (n_{ch} - m_{ch})}$$

and $\delta_{hi}^{r(t)}$ equals 1 if unit i is selected in the replicate group at time point t and zero otherwise;

- calculate weights defined analogously for $i \in s_h^{(2)}$;
- calculate $\tilde{w}_{hi}^{*(r)} = w_{hi}^{*(r)} g_i^{(r)*}$ for $i \in s_h^{(1)}$, $s_h^{(2)}$, where $g_i^{(r)*}$ has the same form as g_i but is calculated using the weights $w_{hi}^{*(r)}$ instead of $w_{hi}^{(r)}$;
- calculate $\hat{\Delta}_{\text{reg}}^* = \hat{Y}_{\text{reg}}^{*(2)} - \hat{Y}_{\text{reg}}^{*(1)}$, where $\hat{Y}_{\text{reg}}^{*(r)} = \sum_{i \in s_h^{(r)}} \tilde{w}_{hi}^{*(r)} y_i$. The WOSB variance estimator is given by

$$\widehat{\text{Var}}_B(\hat{\Delta}_{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^R (\hat{\Delta}_{\text{reg}}^* - \hat{\Delta}_{\text{reg}})^2,$$

$$\text{where } \hat{\Delta}_{\text{reg}} = \hat{Y}_{\text{reg}}^{(1)} - \hat{Y}_{\text{reg}}^{(2)} \text{ and } \hat{Y}_{\text{reg}}^{(r)} = \sum_{i \in s_h^{(r)}} \tilde{w}_{hi}^{(r)} y_i.$$

The proof that $\widehat{\text{Var}}_B(\hat{\Delta}_{\text{reg}})$ is unbiased is straightforward and is similar to the proof that $\text{Var}_B(\hat{\theta})$ is unbiased (see section 4).

The approach described above requires a separate set of replicate weights for movement and level variance estimates. Roberts, Kovačević, Mantel and Phillips (2001) consider approximate Bootstrap variance estimators of movement that only use the level replicate weights, hence reducing computational costs and simplifying the method and its implementation in a computing system.

6. Simulation study

This section summarizes a simulation study for point-in-time and movement estimates carried out to empirically measure the bias and variability of WOSB and WSB over repeated sampling when $R = 100$. A population was generated at time points 1 and 2 from the following models, $y_i^{(1)} = (0.75x_{1i} + 0.25x_{2i}) W(0, 2.5, 1)$ and $y_i^{(2)} = 1.5y_i^{(1)} W(0, 5, 1)$, where the auxiliary variables are given by $x_{1i} = 0.25x_{2i} + 0.75 [100L(0, 1, 1)]$ and $x_{2i} = 100L(0, 1, 1)$ where $W(\mu, \gamma, \alpha)$ and $L(\mu, \gamma, \alpha)$ are the Weibull and Log-normal distributions with location, shape and scale parameters given by μ, γ and α . These distributions reflect the long tails that are typical of economic survey data. The times 1 and 2 populations were of size 3,000, with 2,500 population units common to both time points. Each population unit, i , was assigned to one of 5 strata at both time points using z_i , where $z_i = x_{1i} W(0, 2.5, 1)$ and the stratum boundaries were $z_i = 50, 100, 150, 250$. This resulted in stratum population sizes that ranged from 400 to 1,000.

A total of 3,000 simulated stratified SRSWOR were taken from the population at times 1 and 2, where $n_h^{(1)} = 12$, $n_{ch}^{(1)} = n_{ch}^{(2)} = 4$ and $n_{ch} = 8$ for all h and $t = 1, 2$. For WOSB the replicate sample sizes are given in sections 4 and 5. For WSB the replicate sample sizes for movements were $m_{ch} = [n_{ch} - 1]$, $m_{ch}^{(1)} = [n_{ch}^{(1)} - 1]$ and $m_{ch}^{(2)} = [n_{ch}^{(2)} - 1]$ and for levels were $m_h^{(r)} = n_h^{(r)} - 1$. The WSB estimator for movements has the same form as WOSB but has a slightly different scaling factors and takes replicate samples with replacement.

From each of the 3,000 simulated samples \hat{Y}_{reg}^j is calculated, where \hat{Y}_{reg}^j is given by \hat{Y}_{reg} with $\mathbf{x}_i = (x_{1i}, x_{2i})$, $\sigma_i = 1$ and $j = 1, 2, \dots, 3,000$. The true standard error of \hat{Y}_{reg} is calculated by

$$S = \sqrt{\frac{1}{3,000} \sum_{j=1}^{3,000} (\hat{Y}_{\text{reg}}^j - Y)^2}.$$

The Bootstrap's estimated standard error of \hat{Y}_{reg} from the j^{th} sample is

$$\hat{S}^j = \sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{Y}_{reg}^{j*(r)} - Y)^2},$$

where $\hat{Y}_{reg}^{j*(r)}$ is defined analogously to $\hat{Y}_{reg}^{*(r)}$. The Relative Bias (RB) of the Bootstrap's standard error is

$$RB(\hat{S}) = \frac{1}{3,000\hat{S}} \sum_{j=1}^{3,000} (\hat{S}^j - S).$$

The Relative Root Mean Squared Error (RRMSE) of the Bootstrap's estimated standard error is

$$RRMS(\hat{S}) = \frac{1}{\hat{S}} \sqrt{\frac{1}{3,000} \sum_{j=1}^{3,000} (\hat{S}^j - S)^2}.$$

Similar definitions for RRMSE and bias are used when estimating the movement variance. The 95% coverage probabilities, the percentage of 95% confidence intervals containing the true population total, of WOSB and WSB for levels and movement are also compared.

The results in Table 1 show that the RB and the RRMSE of the WOSB and WSB are both acceptably small. The bias of WSB's time point 1 estimates are slightly higher than WOSB resulting in slightly worse coverage probabilities.

Table 1
Bootstrap estimate of the standard error for movements and point-in-time estimates

Method	Time point 1			Movement		
	RB	RRMSE	C95%	RB	RRMSE	C95(%)
WOSB	0.7	17.3	94.7	2.1	20.7	95.3
WSB	-3.1	15.8	93.7	-1.3	19.6	94.6

7. Summary

From the simulation results, both the WOSB and WSB were considered to be reliably accurate over repeated sampling. Conditional on the sample, the WOSB was found to be significantly more efficient (up to 50%) than WSB for stratified sampling when the stratum sample size is

sometimes small. As a result, the WOSB was implemented in ABSEST.

References

Brewer, K.R.W., Gross, W.F. and Lee, G.F. (1999). PRN sampling: The Australian experience. *Proceedings of the International Associations of Survey Statisticians*, Helsinki.

Estevao, V., Hidiroglou, M.A. and Särndal, C.-E. (1995). Methodological principles for a generalised estimation system at Statistics Canada, *Journal of Official Statistics*, 11, 2, 181-204.

Kovar, J., Rao, J.N.K. and Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 25-44.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 4, 363-378.

Preston, J., and Chipperfield, J.O. (2002). Using a generalised estimation methodology for ABS business surveys. Methodology Advisory Committee (available at www.abs.gov.au).

Rao, J.N.K., and Wu, C.F.J. (1984). Bootstrap inference for sample surveys. *Proceeding Section on Survey Methods Research, Journal of the American Statistical Association*, 106-112.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, March, 83, 401, 231-241.

Roberts, G., Kovačević, M., Mantel, H. et Phillips, O. (2001). Cross sectional inference based on longitudinal surveys: Some experience with Statistics Canada Surveys. Presented at the 2001 FCSM Research Conference, Arlington, VA, November 14-16, proceedings available at www.fcsm.gov.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

Tam, S.M. (1985). On covariance in finite population sampling. *The Statistician*, 34, 429-433.

Yung, W., and Rao, J.N.K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, 22, 23-31.

Bayesian estimation in small areas when the sampling design strata differ from the study domains

Jacob J. Oleson, Chong Z. He, Dongchu Sun and Steven L. Sheriff¹

Abstract

The purpose of this work is to obtain reliable estimates in study domains when there are potentially very small sample sizes and the sampling design stratum differs from the study domain. The population sizes are unknown as well for both the study domain and the sampling design stratum. In calculating parameter estimates in the study domains, a random sample size is often necessary. We propose a new family of generalized linear mixed models with correlated random effects when there is more than one unknown parameter. The proposed model will estimate both the population size and the parameter of interest. General formulae for full conditional distributions required for Markov chain Monte Carlo (MCMC) simulations are given for this framework. Equations for Bayesian estimation and prediction at the study domains are also given. We apply the 1998 Missouri Turkey Hunting Survey, which stratified samples based on the hunter's place of residence and we require estimates at the domain level, defined as the county in which the turkey hunter actually hunted.

Key Words: Hierarchical Bayes; Markov chain Monte Carlo; Bayesian prediction; Random sample size; Spatial correlation; Stratification.

1. Introduction

Small sample sizes occur often when analyzing sample survey data. These small sample sizes arise frequently when studying subpopulations such as socio-demographic groups. We may also consider spatial regions and time periods as subpopulations or study domains. Due to small sample sizes, direct survey estimators could be highly unreliable. Estimation stemming from areas with small sample sizes has been termed small area estimation (SAE). Rao (2003) gives a nice review of many SAE techniques. Some recent small area review papers are found in Rao (2005) and Jiang and Lahiri (2006).

Appropriate models are needed in order to produce reliable small area statistics. Different model-based methods include the empirical best prediction method (see Prasad and Rao 1990, Jiang, Lahiri and Wan 2002, Das, Jiang and Rao 2004, Jiang and Lahiri 2006) and Bayesian methods (see Malec, Sedransk, Moriarity and LeClere 1997, Ghosh, Natarajan, Stroud and Carlin, 1998, He and Sun 2000). For a good review on Bayesian small area estimation, the readers are referred to Rao (2003). This paper concerns a practical implementation of Bayesian methodology. One critical step in implementing a Bayesian model is selecting prior distributions. The propriety of the posterior distribution and the robustness of the priors should be carefully examined. When an MCMC simulation such as Gibbs sampler is used in the computation, the convergence of the Gibbs chain must be monitored. For more details, see Carlin and Louis (2000).

We consider a stratified cluster sampling design where within each stratum clusters are selected by simple random sampling without replacement. In our application, clusters are of unequal sizes and the cluster sizes are unknown at the time of designing the survey. We consider a domain estimation problem where domains cut across the design clusters. As a result, domain population size is unknown and domain sample size is random. In our application, realized sample sizes for the domains are small and as such standard design-based domain estimation techniques (see Cochran 1977, Lohr 1999) are unreliable. We propose a fully Bayesian hierarchical model to get around the problem.

We begin by obtaining estimates of success rates and population sizes simultaneously at the small area level for individuals from each of the design strata. The estimates are obtained by borrowing information from the neighboring small areas. This is done through a spatial structure built into the Bayesian model. Therefore, the resulting estimates are much more stable than the direct survey methods. We then compute a weighted average of the success rates from the design stratum for the final small area estimate. For example, if a county is the small area, we compute a county-specific success rate estimate for each of the design strata and average them for a single county estimate. To combine the sampling design strata, the individual stratum population sizes should be known. In the case where these population sizes are not known, they can be estimated using our proposed model. This work is motivated by applying Bayesian methods in estimating the turkey hunting success rates at the county level in Missouri. We propose a new

1. Jacob J. Oleson, Department of Biostatistics C22-GH, The University of Iowa, Iowa City, Iowa, 52242-1009, U.S.A.; Chong Z. He and Dongchu Sun, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, Missouri 65211-6100, U.S.A.; Steven L. Sheriff, Resource Science Center, Missouri Conservation Department, 1110 South College Avenue, Columbia, Missouri 65201.

family of generalized linear mixed models with correlated random effects when there is more than one unknown parameter. We will call this a bivariate generalized linear mixed model (bivariate GLMM). A generalized linear mixed model (GLMM) with possible correlated random effects is often used when there is only one unknown parameter (Sun, Speckman and Tsutakawa 2000). The proposed model estimates both the population size and the parameter of interest and, hence, advances the current state of small area research.

The above-described scenario is present in the 1998 Missouri Turkey Hunting Survey (MTHS). This was a spring postseason mail survey that provided the Missouri Department of Conservation with information concerning the number of turkeys harvested by hunters on each day of the hunting season and in which county the harvest occurred. Also, the total number of trips made to the counties by these hunters on each hunting day was recorded. Hunting success rates were then calculated from this information.

The MTHS example is presented in detail in Section 2. It is followed in Section 3 by a summary of the proposed methodology and by general formulae. Although we estimate success rates, the methodology is generalizable. We also give general formulae to find the estimates and predictions for the small areas as well as full conditional distributions for use with MCMC simulations. Final comments are given in Section 4.

2. 1998 Missouri turkey hunting survey

2.1 Background to 1998 MTHS

The Missouri Department of Conservation began biennially in 1986 to track hunter tendencies with the MTHS. This survey asked the hunter what county he/she hunted in, on what day that occurred, and if the hunt was successful or not. It began as a simple random sample of all spring turkey hunting permit holders. He and Sun (1998) used the information from the 1996 survey to estimate turkey hunter success rates in all 114 counties of Missouri with a Bayesian Beta-Binomial model. He and Sun (2000) estimated county-specific hunting success rates per week of the hunting season. Only one harvested turkey was allowed per week in 1996 spring hunting season. They used a GLMM and estimated only success rates. Oleson and He (2004) extended this model in order to estimate hunting success rates for each day of the hunting season. They found significant auto-correlation among the days of the hunting season and among the counties of Missouri when estimating success rates.

In 1998, the MTHS sampling scheme was changed. The frame is still the list of all turkey hunters registered to hunt in Missouri and contains, among other things, information on each hunter's county of residence. Simple random samples used previously put too much weight on the heavy population masses of Kansas City and St. Louis. Hence, counties near these metropolises received large samples and counties further away (*e.g.*, the southern tiers) received insufficient samples. Ideally we would like to stratify by the county where hunters pursued turkeys to draw samples representative at this domain level. Information about where the hunters pursued turkeys is unavailable until after questionnaires are returned, meaning this type of stratification is not possible.

An alternative is to stratify by where the hunter lives since hunters tend to hunt near where they live (in locations with which they are familiar). This causes a problem in estimating the parameters of interest which are hunting success rates per county. We would like estimates based on hunting location, but the sampling design uses the hunter's place of residence. The new sampling design of MTHS is a stratified simple random sampling of clusters with unequal sizes. In this case, a cluster represents a registered hunter and its elements are the hunting trips for that hunter. The hunter's place of residence is used as a stratification factor. The design strata are: 1) Non-residents of Missouri, 2) Residents of Northern Missouri, 3) Residents of Southern Missouri, 4) Residents of St. Louis metro area, and 5) Residents of Kansas City metro area. Figure 1 shows the boundaries that were used in determining the four Missouri resident strata. These are based on the first three digits of the postal zip code. Proportional allocation was used to determine the number of sampled hunters in each stratum. As shown in Table 1, there were 110,691 total permit holders. A sample of 8,000 was proportionally allocated to each of the five strata. In Table 1 the column "% of Sample" refers to the percent of the overall sample that comes from that particular stratum (sample replied/total sample replied). The column "% of Strata Sampled" contains the percent of hunters who were sampled from that stratum (sample replied/total permits). The number of clusters in the population and the number of clusters in the sample for each stratum are known fixed numbers. The hunting trips taken by each hunter are the elements within the cluster. Throughout the remainder of the paper, population size is the number of hunting trips taken by all hunters in the frame, termed hunting pressure. For the MTHS example, the population size is unknown but not random; the sample size is best considered random since we do not know how many trips each hunter will take. Note that a hunter does not have to hunt in his/her design stratum. The hunter may hunt in many different counties during the season. For each study

domain (*i.e.*, county), the population size is again unknown but not random; the sample size is again random where we do not know in which county a hunter is going to hunt. Small sample sizes can still result; for example, Dunklin County, Pemiscot County, and New Madrid County in southeast Missouri had zero trips reported to them throughout the 3-week season in 1998 and Lawrence County in southwest Missouri had zero trips on the first day and only five the remainder of the first week.

Table 1 Sample sizes and response rates from 5 strata for 1998 MTHS

Design Stratum	Total Permits	Sample Sent	Sample Replied	Response Rate (%)	% of Sample	% of Strata Sampled
Non-Resident	19,798	1,600	1,180	73.75	22.2	5.96
North Missouri	21,756	1,532	975	63.64	18.3	4.48
South Missouri	34,375	2,421	1,509	62.33	28.4	4.39
St. Louis	19,959	1,405	969	68.97	18.2	4.85
Kansas City	14,803	1,042	688	66.03	12.9	4.65
Total	110,691	8,000	5,321	66.51	100	4.81

The MTHS is a post-season mail survey where the questionnaires are mailed to the hunter's listed residence. The survey has questions specifying each day of the hunting season and asking hunters whether or not they hunted, in what county they hunted, and if their hunt was successful or not. In total 41% of the residents responded to the first mailing. Those who did not respond within two months

were mailed the survey again, to which 26% responded from the second mailing. The survey was mailed a third time to those who had still not responded after two months with 20% of the third mailing responding. This resulted in an overall response rate of 66.5%. From Table 1 we see that non-residents of Missouri had the highest response rate at nearly 75%. The two metro areas of St. Louis and Kansas City then followed with response rates of 69% and 66%, respectively. The rural areas in northern Missouri and southern Missouri had the lowest response rates at approximately 63%.

The 1998 spring turkey hunting season in Missouri consisted of the 21 consecutive days beginning on Monday, 19 April. During the first week, a turkey hunter could harvest one bearded turkey. If successful during the first week, the hunter was allowed to harvest only one additional bearded turkey during the last two weeks of the season. If the hunter was unsuccessful during the first week, the hunter could harvest two bearded turkeys during the second two weeks, but only one bird per day could be taken. During this 3-week season, the turkey biologist wanted information concerning hunter success for the first day due to the opening day effect in the hunting season when hunting pressure normally exceeds that of any other day of the season. The remaining six days of the first week are combined into a second time period of interest, called week 1. Week 2 and week 3 are modeled separately to give a total of four time periods.

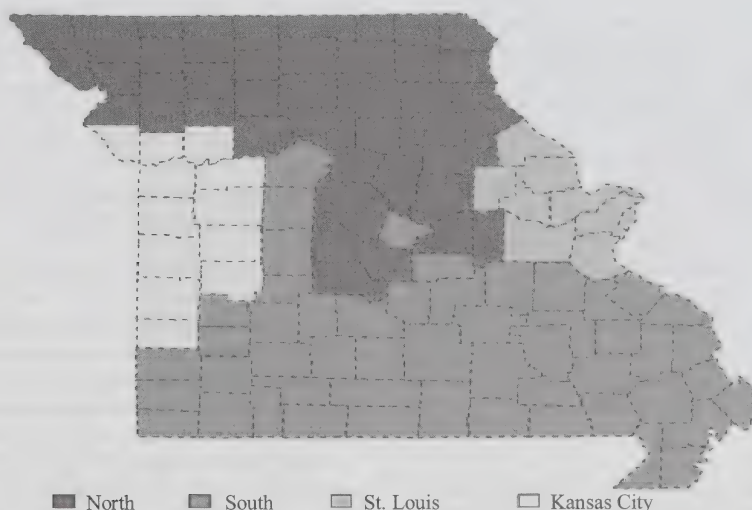


Figure 1 Sampling domain boundaries for the four Missouri resident sampling domains

The small area of interest is the county-time period combination for turkey hunters from each stratum. Turkey population management focuses on the county level as the smallest areal unit of interest due to hunters' ability to indicate where they have hunted in reporting their success. A hunting license is required to hunt, thus we know how many total hunters there are in Missouri and in which stratum they live. Spreading the data out across 114 counties produces very sparse data within each of the counties. For this reason, rather than looking at success rates on a daily basis, we will use four time periods as defined above. In calculating the county-specific success rates, however, we use the number of hunting trips in each county. We do not know how many trips each hunter will take or in which county they will hunt. Thus, while we know the number of hunters in each stratum, the number of trips is unknown. Furthermore, while we know the number of hunters who have been sampled from each stratum, the number of trips to a specific county is unknown and random because a different sample of hunters will yield a different number of trips. There are two principle parameters to estimate and one random variable to predict. The first parameter is the total number of hunting trips taken by all hunters, known as hunting pressure. This is important to wildlife managers who are concerned about the quality of the hunting experience. Too many hunters in an area have the tendency to interfere with each other's hunt, which in turn lowers the quality of the hunting experience. The second parameter to estimate is the hunting success rate, or the proportion of turkeys harvested per hunting trip. If there are not enough turkeys in a county, wildlife managers will close the county for a number of years in order for the turkey population to recover. The random variable to predict is the total number of turkeys harvested. In Missouri, every turkey that is harvested must be reported. In 1998, turkey hunters were required to go to a check station where their turkeys were tallied. It was expensive to maintain these check stations. One of purposes of this project is to predict the total number of turkeys harvested at the county level and compare this to the actual number recorded at the check stations.

Next we present a model to estimate hunting success rate and hunting pressure, as well as predict total turkey harvest simultaneously. The model accounts for random sample sizes that the previous models of He and Sun (1998), He and Sun (2000), and Oleson and He (2004) did not.

2.2 The Model

Let p_{ijk} denote the success rate, y_{ijk} the number of successes, n_{ijk} the number of trips in the sample, and N_{ijk} the total number of trips, in county i , at time j , for individuals from stratum k . We model turkey harvests with independent binomial distributions

$$(y_{ijk} | n_{ijk}, p_{ijk}) \stackrel{\text{ind}}{\sim} \text{Binomial}(n_{ijk}, p_{ijk}), \quad (1)$$

where, $i = 1, \dots, I$ is the county (*i.e.*, study domain), $j = 1, \dots, J$ is the time period, and $k = 1, \dots, K$ is the design stratum. Here $I = 114$, $J = 4$, $K = 5$ for MTHS. The previous analyses of the MTHS assumed a fixed sample size, n_{ijk} , which we believe is best considered random. We don't know the number of trips made to each county until after the survey has been collected. We also don't know the total number of hunting trips taken in the small area, N_{ijk} , only the number of potential hunters. Since hunters must stop after their second turkey, counties with higher success rates would be expected to have fewer hunting trips (or days) for the same number of harvested birds than a county with lower success rates. If there is a correlation, then the sample size must be considered random. Also, in Bayesian hierarchical modeling, if the distribution of a sample size is independent of the distribution of the response variable, then the estimates are identical for random and non-random (fixed) sample size n , see Durbin (1969). The estimated success rate is smoother for a fixed n_{ijk} than when it is random (Woodard, He and Sun 2003). Malec *et al.* (1997) applied the Bayesian small area estimation to the National Health Interview Survey. There are two major differences between our model and that of Malec *et al.* (1997). First, the population sizes, N_{ijk} , are known in their model but unknown in our model. This is the main reason that we introduce the bivariate GLMM. Secondly, the logit of the success rate is modeled as a linear function of covariates in their model. Therefore, the estimates depend on the values of covariates but not the spatial locations. We will add a spatial component to the logit of success rates in addition to the covariates so that the estimates depend on both covariates and spatial locations. This is necessary if some important covariates are not available.

To incorporate the randomness of the sample sizes, we model n_{ijk} with Poisson distributions

$$(n_{ijk} | N_{ijk}) \stackrel{\text{ind}}{\sim} \text{Poisson}(R_k N_{ijk}). \quad (2)$$

The mean and variance of the Poisson distribution for n_{ijk} is a constant multiplied by the population size, N_{ijk} . This constant R_k is the ratio of the number of hunters in the sample for stratum k to the total number of hunters from stratum k . This ratio can be calculated from Table 1.

For the Poisson distribution, the overall sample size, n_k is considered random. We presented in the previous section why we consider this assumption to be appropriate. If n_k were fixed, then the multinomial distribution would be a more appropriate model. The likelihoods of these two

approaches are very similar and yield comparable results in either context (see Agresti 2002, pages 8-9).

We model p_{ijk} using its logit function and N_{ijk} using a logarithm transformation, i.e.,

$$\eta_{ijk} = \log\left(\frac{p_{ijk}}{1 - p_{ijk}}\right), \quad \omega_{ijk} = \log(N_{ijk}).$$

Linear mixed models are used for the priors on both η_{ijk} and ω_{ijk} by assuming

$$\eta_{ijk} = \theta_{1jk} + u_{1ik} + e_{1ijk},$$

$$\omega_{ijk} = \theta_{2jk} + u_{2ik} + e_{2ijk}.$$

Here for $a = 1, 2$, θ_{ajk} denotes fixed effect due to the j^{th} time in stratum k , u_{aik} represents a random county effect, and random errors e_{ajik} are iid $N(0, \delta_{ak}^{(e)})$.

To complete the Bayesian hierarchical model, we need to specify the priors for $\theta_{ak} = (\theta_{a1k}, \dots, \theta_{aJk})'$, $\mathbf{u}_{ak} = (u_{a1k}, \dots, u_{aIk})'$, and $\delta_{ak}^{(e)}$. He and Sun (2000) and Oleson and He (2004) show that there is significant spatial correlation among counties of Missouri in estimating the success rates. They use a conditional auto-regressive (CAR) structure to model spatial dependence between neighboring counties. The joint density of \mathbf{u}_{ak} is given by

$$f(\mathbf{u}_{ak}) =$$

$$(2\pi\delta_{ak}^{(u)})^{-\frac{I}{2}} |\mathbf{I} - \rho_{ak}\mathbf{C}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2\delta_{ak}^{(u)}} \mathbf{u}_{ak}' (\mathbf{I} - \rho_{ak}\mathbf{C}) \mathbf{u}_{ak}\right\}. \quad (3)$$

Here $\delta_{1k}^{(u)}$ and $\delta_{2k}^{(u)} > 0$ are variance components and $\mathbf{I} - \rho_{ak}\mathbf{C}$ is a nonnegative definite symmetric matrix (Besag 1974). \mathbf{I} is the $I \times I$ identity matrix and \mathbf{C} is an adjacency matrix whose component $\{c_{ij}\}$ is 1 if areas i and j share a common boundary with $\{c_{ij}\} = 0$. We also define ρ_{1k} and ρ_{2k} to be the spatial correlation parameters. Let $\lambda_1 \leq \dots \leq \lambda_I$ be the eigenvalues of the matrix \mathbf{C} . These matrices, $\mathbf{I} - \rho_{ak}\mathbf{C}$, are positive definite if $\lambda_1^{-1} \leq \rho_{ak} \leq \lambda_I^{-1}$ (Clayton and Kaldor 1987). For the Missouri data, $I = 114$ and the numerical values of λ_1 and λ_{114} are -2.8931 and 5.6938, respectively. This means that the density of \mathbf{u}_{ak} exists if ρ_{ak} is in (-0.3457, 0.1756).

For the remaining priors, we assume the following. θ_{ajk} is normal with a mean μ_{ajk} and variance τ_{ajk} . Let ρ_{ak} be uniformly distributed on the interval $(\lambda_1^{-1}, \lambda_I^{-1})$. Finally, a common prior for the variance components $\delta_{ak}^{(e)}$ and $\delta_{ak}^{(u)}$ is inverse gamma (Gelman, Carlin, Stern and Rubin 1995) whose densities are proportional to

$$\frac{1}{\delta_{ak}^{(e)\alpha_{ak}^{(e)}+1}} \exp\left(-\frac{\beta_{ak}^{(e)}}{\delta_{ak}^{(e)}}\right) \text{ and } \frac{1}{\delta_{ak}^{(u)\alpha_{ak}^{(u)}+1}} \exp\left(-\frac{\beta_{ak}^{(u)}}{\delta_{ak}^{(u)}}\right),$$

respectively.

To evaluate the posterior distribution, we apply MCMC methods such as Gibbs sampling (Gelfand and Smith 1990) to obtain samples from the posterior distribution. For an

overview of MCMC methodologies see Gelman *et al.* (1995), Gilks, Richardson and Spiegelhalter (1996) and Robert and Casella (1999). The full conditional distributions required for this evaluation may be found directly from those given in Fact 1 in Section 3. Most of these conditional distributions (θ_{ak} , \mathbf{u}_{ak} , $\delta_{ak}^{(e)}$, $\delta_{ak}^{(u)}$) are of standard forms and are easily sampled from. Other full conditional distributions (η_{ijk} , ω_{ijk} and ρ_{ak}) have log-concave densities for the MTHS. The adaptive rejection sampling method of Gilks and Wild (1992) can be used to generate random samples from log-concave densities.

2.3 Bayesian estimation and prediction

At this point, we have obtained estimates of (p_{ijk}, N_{ijk}) . We wish to pool the estimates from the stratum together in order to estimate p_{ij} and N_{ij} . We will also predict the unobserved number of harvests in county i , denoted h_i .

To obtain the estimates of (p_{ij}, N_{ij}) , let $(p_{ijk}^{(l)}, N_{ijk}^{(l)})$, $l = 1, \dots, L$ be the output from the sampled Gibbs chain after the burn-in sample. Define

$$p_{ij}^{(l)} = \frac{\sum_{k=1}^K p_{ijk}^{(l)} N_{ijk}^{(l)}}{\sum_{k=1}^K N_{ijk}^{(l)}}, \quad l = 1, \dots, L.$$

The posterior mean and variance of p_{ij} can then be approximated by

$$\hat{E}(p_{ij}) = \frac{1}{L} \sum_{l=1}^L p_{ij}^{(l)} \quad (4)$$

and

$$\hat{V}(p_{ij}) = \frac{1}{L-1} \sum_{l=1}^L \{p_{ij}^{(l)}\}^2 - \frac{L}{L-1} \{\hat{E}(p_{ij})\}^2, \quad (5)$$

respectively. Similarly define $N_{ij}^{(l)} = \sum_{k=1}^K N_{ijk}^{(l)}$. The posterior mean and variance of N_{ij} can be approximated from the MCMC simulation output as well.

We now focus on Bayesian prediction of the unobserved harvest by their posterior predictive distributions. We have that y_{ijk} represents the number of harvested turkeys in county i , of time period j , and of sampling stratum k for those in the sample, whereas for those not in the sample, we let the number of harvested turkeys in county i , during time j , stratum k be represented as y_{ijk}^* . Thus, the number of turkeys harvested in county i at time j for hunters from stratum k is $h_{ijk} = y_{ijk} + y_{ijk}^*$. Here y_{ijk} is a known value and we need only find y_{ijk}^* . We may think of y_{ijk}^* given $(n_{ijk}, N_{ijk}, p_{ijk})$ as a binomial random variable in the form of (1). Thus

$$(y_{ijk}^* | n_{ijk}, N_{ijk}, p_{ijk}) \stackrel{\text{ind}}{\sim} \text{Binomial}(N_{ijk} - n_{ijk}, p_{ijk}). \quad (6)$$

Let $(p_{ijk}^{(l)}, N_{ijk}^{(l)})$, $l = 1, \dots, L$, be the output from running a Gibbs chain after a burn-in sample. The predictive mean of y_{ijk}^* given data $\mathbf{d} = \{(y_{ijk}, n_{ijk}) : i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K\}$ is then

$$\hat{E}(y_{ijk}^* | \mathbf{d}) = \frac{1}{L} \sum_{l=1}^L (N_{ijk}^{(l)} - n_{ijk}) P_{ijk}^{(l)}. \tag{7}$$

Finally we write

$$\begin{aligned} \hat{h}_{ijk} &= y_{ijk} + \hat{E}(y_{ijk}^* | \mathbf{d}), \\ \hat{h}_i &= \sum_{j=1}^J \sum_{k=1}^K \hat{h}_{ijk}. \end{aligned} \tag{8}$$

The variability of h_i is given by

$$\begin{aligned} \hat{V}(h_i | \mathbf{d}) &= \frac{1}{L-1} \sum_{l=1}^L \sum_{j=1}^J \sum_{k=1}^K (N_{ijk}^{(l)} - n_{ijk}) P_{ijk}^{(l)} (1 - P_{ijk}^{(l)}) \\ &\quad + \frac{1}{L-1} \sum_{l=1}^L \left\{ \sum_{j=1}^J \sum_{k=1}^K [y_{ijk} + (N_{ijk}^{(l)} - n_{ijk}) P_{ijk}^{(l)}] - \hat{h}_i \right\}^2. \end{aligned}$$

The proof is a straightforward application of conditional expectations. Note that h_{ijk} is defined as a random variable. It does not equal $N_{ijk}P_{ijk}$ just as y_{ijk} does not equal $n_{ijk}P_{ijk}$. Therefore, one should not simply use $\hat{N}_{ijk}\hat{P}_{ijk}$ to estimate h_{ijk} although it might be very close if \hat{N}_{ijk} is large enough.

2.4 Model fitting

Fifteen-thousand iterations were run for the Gibbs chain, of which 5,000 were discarded as a burn-in sequence. Thus, posterior estimates are based on 10,000 autocorrelated samples from the posterior distribution. In monitoring the convergence of Gibbs sampling, we have used the diagnostics of Heidelberger and Welch (1983) as well as graphical monitoring of the sample paths. Convergence diagnostics and posterior summaries were performed with the BOA software (Smith 2005). We let variance components $\delta_{ak}^{(e)}$ and $\delta_{ak}^{(u)}$ follow a non-informative IG(2,1) prior which gives a mean of one and infinite variance. We give data-dependent priors to $\theta_{1jk} \sim N(-1.5, 16)$ and $\theta_{2jk} \sim N(4, 25)$. To obtain a data-dependent prior, we modeled the following two steps because a completely arbitrary data-dependent prior might lead to unreliable posterior estimates (Wasserman 2000). First, we began with noninformative (but conjugate) priors to obtain the posterior means and standard deviations for each of these parameters through MCMC simulations. We then set data-dependent prior means to be close to the posterior means of θ_{1jk} and θ_{2jk} and the data-dependent prior standard deviations to be about ten times the posterior standard deviations, after the posterior estimates were obtained using the noninformative priors. The estimates using these two approaches were quite similar, but the model using the data-dependent priors gave smaller variances.

We fit many simplified models for comparison and model checking. As a model selection tool we use the Deviance Information Criterion (DIC) suggested by Spiegelhalter, Best, Carlin and van der Linde (2002). DIC is a generalization of the Akaike Information Criterion that

measures the deviance over an MCMC run and includes a penalized fit measure taking into account the model dimension. Smaller values of DIC indicate a better-fitting model. Our model with data-dependent priors had a DIC = 13,910.5. No alternate models had a significantly reduced DIC value from this model. The model with non-informative priors on θ_{1jk} and θ_{2jk} had DIC = 14,700.4. A reduced model with common correlation parameters $\rho_{11} = \rho_{12} = \rho_{13} = \rho_{14} = \rho_{15}$ and $\rho_{21} = \rho_{22} = \rho_{23} = \rho_{24} = \rho_{25}$ had DIC = 13,896.9.

As another model check, we have calculated statewide averages of the estimates using both the simple naive design-based estimates and Bayesian estimates which follow in Table 2. At the statewide level, sample sizes are large enough to consider the design-based estimates reliable. The statewide design-based estimates and model estimates match closely. Thus, the Bayes estimator performs well in terms of the design consistency property (see You and Rao 2003 and Jiang and Lahiri 2006). We note that the first day estimates are slightly lower due to smoothing, but the success rate estimates are still much higher for the first day than any other time period.

Table 2 Success rate estimates for 1998 MTHS					
Design Stratum	Period	Kills	Trips	Design-based	Bayesian
Non-Resident	Day 1	105	424	0.248	0.236
	Week 1	271	1,759	0.154	0.155
	Week 2	198	1,481	0.134	0.132
North	Week 3	75	635	0.118	0.120
	Day 1	82	452	0.181	0.171
	Week 1	211	1,535	0.138	0.137
South	Week 2	161	1,663	0.097	0.097
	Week 3	99	1,129	0.088	0.090
	Day 1	138	751	0.184	0.178
St. Louis	Week 1	224	2,475	0.091	0.092
	Week 2	186	2,375	0.078	0.069
	Week 3	90	1,748	0.052	0.053
Kansas City	Day 1	54	346	0.156	0.149
	Week 1	91	1,280	0.071	0.073
	Week 2	85	1,257	0.068	0.069
Overall	Week 3	45	828	0.054	0.059
	Day 1	52	262	0.199	0.181
	Week 1	95	894	0.106	0.108
Total	Week 2	92	942	0.098	0.099
	Week 3	55	611	0.090	0.093
	Day 1	431	2,235	0.193	0.181
Overall	Week 1	892	7,943	0.112	0.111
	Week 2	722	7,718	0.094	0.088
	Week 3	364	4,951	0.074	0.075
Total		2,409	22,847	0.105	0.102

2.5 Data analysis for MTHS

Posterior means and standard deviations for parameters under all design strata are listed in Table 3. We note that the mean and variance estimates of θ_{12k} and θ_{13k} as well as θ_{22k} and θ_{23k} for $k = 1, \dots, 5$ are approximately equal. This gives reason to believe that success rate estimates and hunting pressure estimates are similar for week 1 and 2 of the hunting season. There remains a difference in the first day and week 3 of the hunting season, though (Vangilder, Sheriff and Olsen 1990, Kimmel 2001).

Table 3 Posterior means (Standard deviations) of model parameters

	Non-Res	North	South	St. Louis	K.C.
$\hat{\delta}_{e1}$	0.133 (0.034)	0.117 (0.030)	0.136 (0.036)	0.164 (0.051)	0.143 (0.040)
$\hat{\delta}_{e2}$	0.096 (0.017)	0.044 (0.007)	0.041 (0.006)	0.061 (0.011)	0.060 (0.011)
$\hat{\delta}_{z1}$	0.146 (0.040)	0.148 (0.041)	0.140 (0.038)	0.166 (0.053)	0.186 (0.059)
$\hat{\delta}_{z2}$	0.570 (0.092)	0.511 (0.081)	0.633 (0.094)	0.373 (0.062)	0.356 (0.063)
$\hat{\theta}_{11}$	-1.209 (0.131)	-1.530 (0.130)	-1.244 (0.135)	-1.420 (0.159)	-1.472 (0.160)
$\hat{\theta}_{12}$	-1.613 (0.109)	-1.748 (0.111)	-1.867 (0.130)	-1.983 (0.147)	-1.928 (0.150)
$\hat{\theta}_{13}$	-1.801 (0.113)	-2.081 (0.113)	-2.152 (0.129)	-2.032 (0.148)	-2.007 (0.148)
$\hat{\theta}_{14}$	-1.830 (0.132)	-2.135 (0.123)	-2.354 (0.138)	-2.115 (0.156)	-2.008 (0.153)
$\hat{\theta}_{21}$	3.368 (0.098)	3.333 (0.098)	3.297 (0.087)	3.305 (0.088)	3.308 (0.098)
$\hat{\theta}_{22}$	4.582 (0.097)	4.361 (0.093)	4.371 (0.084)	4.349 (0.081)	4.238 (0.090)
$\hat{\theta}_{23}$	4.404 (0.097)	4.429 (0.093)	4.459 (0.088)	4.293 (0.083)	4.229 (0.093)
$\hat{\theta}_{24}$	3.693 (0.101)	4.093 (0.095)	4.039 (0.086)	3.934 (0.082)	3.891 (0.094)
$\hat{\rho}_1$	0.153 (0.019)	0.103 (0.071)	0.163 (0.011)	0.158 (0.016)	0.125 (0.068)
$\hat{\rho}_2$	0.168 (0.007)	0.172 (0.003)	0.171 (0.004)	0.172 (0.004)	0.172 (0.004)

We also point out the correlation parameter estimates of ρ_{1k} and ρ_{2k} . The estimates for ρ_{2k} for the hunting pressure are all near their upper boundary, suggesting a strong relationship between the counties when estimating the number of hunting trips taken. Most of the estimates for ρ_{1k} are more than two standard deviations away from zero as well.

The simple naive design-based success rate estimates for each county are in the first row of Figure 2. Design-based estimates in the counties range from 0 to 0.55 with some

counties having no estimate because $n_{ij} = 0$. Alternatively, the Bayesian estimates for the success rate in county i for time j are plotted in the second row of Figure 2. The Bayesian model success rate estimates produce a much more sensible range from 0.03 to 0.30. We don't expect any county to have a success rate estimate of 0. Also, turkeys are quite difficult to hunt and a high success rate is not sensible (He and Sun 1998, He and Sun 2000, Oleson and He 2004). Thus a lower value, such as that produced from the model, makes more intuitive sense. Standard deviations for the model success rate estimates are given in the third row of Figure 2.

Success rate estimates tend to decrease over the course of the hunting season. In addition, the highest success rate estimates are in the northern portion of Missouri. This was also shown to be true in previous analyses of the MTHS using 1996 data (He and Sun 2000, Oleson and He 2004). We note that the highest success rate estimates are not in the same counties, though. The highest estimated rates from 1998 have moved slightly to the east of where they were in 1996. This shift is due to a time trend as noted by conservationists.

We also produce estimates of the population size N_{ijk} . The model-based estimates of N_{ij} are plotted in the first row of Figure 3. The standard errors for the model-based estimates are plotted in the second row of Figure 3. The values plotted for weeks 1, 2, and 3 are the daily averages for those weeks. It is apparent that more people were hunting on the first day than any other day of the hunting season.

We plot the actual check station data as the first map in Figure 4. Using formula (8) we predict the model-based total harvest h_i in the second map of the first row. From these figures, the model-based predictions look to be reasonably accurate. We expect to see a small amount of overprediction as is shown by comparing the plots. Some hunters may underestimate how often they went out to hunt and overestimate the number of birds they harvested. The overcount may possibly also be attributed to turkeys harvested but not reported at a check station. One more explanation could be that the hunters returning the survey are those who were more successful, and those not returning the survey were less successful or did not hunt. Comparing the predictions from the model to the actual harvest numbers is a confirmation that the model is appropriate and gives accurate predictions. Also, many states do not have check stations and this shows that they could obtain appropriate estimates at the domain level even with a small sample at the domain level.



Figure 2 Hunting success rates from stratified 1998 MTHS. Row 1 - design based estimates of success rates. Row 2 - Bayesian estimates of success rates. Row 3 - Standard deviations of Bayesian estimates of success rates

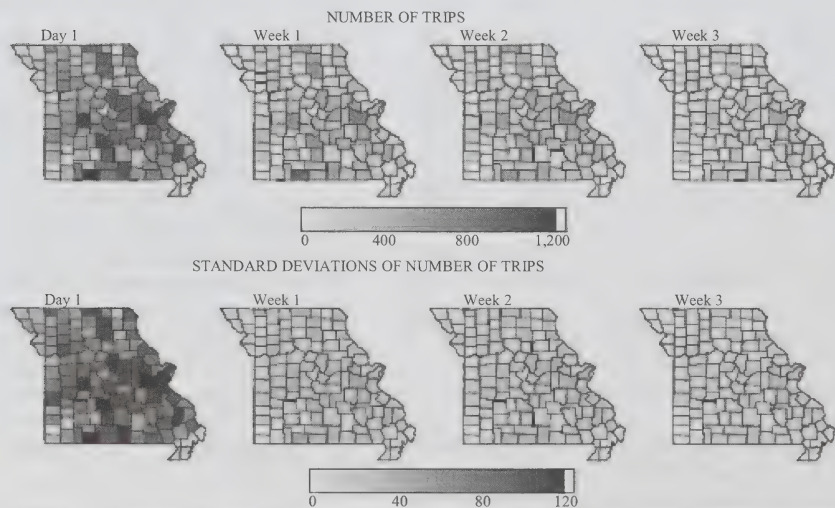


Figure 3 Estimated number of hunting trips from stratified 1998 MTHS. Row 1 - Bayesian estimated number of tips. Row 2 - Standard deviation of Bayesian estimates of number of trips

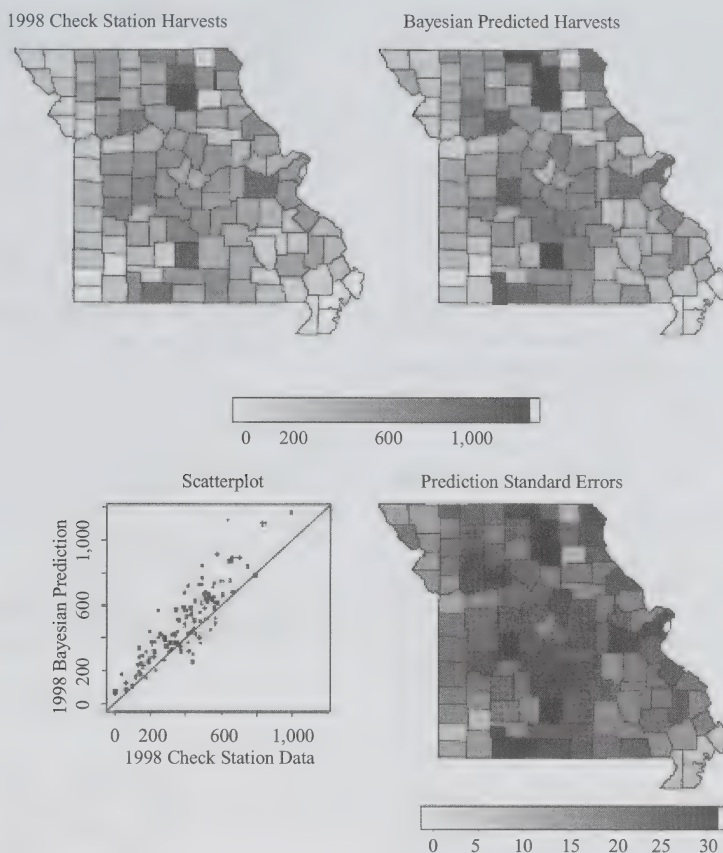


Figure 4 1998 check station data, bayesian prediction, scatterplot of check station by prediction, standard error of bayesian prediction

3. General formulae for a bivariate GLMM

3.1 A general model

We ignore the time component in the general model for simplicity. Let n_{ik} and Y_{ik} be the sample sizes and response variables of interest in study domain i and design stratum k , respectively. We assume that $\{Y_{ik}, n_{ik}\}$ given unknown parameters $\{\eta_{ik}, \omega_{ik}, \phi_1, \phi_2\}$, with $i = 1, \dots, I$, and $k = 1, \dots, K$, are independent. Assume that the conditional density (mass) function of Y_{ik} given n_{ik} belongs to the following family of probability densities,

$$g_1(y_{ik} | \eta_{ik}, n_{ik}, \phi_1) = \exp[A_1(\phi_1)\{y_{ik}\eta_{ik} - B_1(\eta_{ik}, n_{ik})\} + C_1(y_{ik}, n_{ik}, \phi_1)], \quad (9)$$

where η_{ik} is an unknown parameter, and it is often assumed the scale parameter ϕ_1 is known. The probability density function of n_{ik} is in the family,

$$g_2(n_{ik} | \omega_{ik}, \phi_2) = \exp[A_2(\phi_2)\{n_{ik}\omega_{ik} - B_2(\omega_{ik})\} + C_2(n_{ik}, \phi_2)], \quad (10)$$

where ω_{ik} is an unknown parameter equal to a function of the population size N_{ik} , and ϕ_2 is often known. The joint density of Y_{ik} and n_{ik} , a bivariate GLMM, is then

$$p(y_{ik}, n_{ik} | \eta_{ik}, \omega_{ik}, \phi_1, \phi_2) = g_1(y_{ik} | \eta_{ik}, n_{ik}, \phi_1)g_2(n_{ik} | \omega_{ik}, \phi_2). \quad (11)$$

The distribution family (10) is often called a generalized linear model, which includes binomial, Poisson, normal and

gamma distributions. See, for example, Sun *et al.* (2000). The distribution family (9) is a generalization of such a generalized linear model by including an additional parameter. Four special cases of (9) are the binomial, Poisson, normal and gamma distributions which are all a part of the exponential family.

The bivariate GLMM is applicable when estimates at the study domains are of interest and the sample size n_{ik} is considered random as well as being part of the observation. The bivariate GLMM is also useful when estimates of N_{ik} are required.

A linear mixed model can be used as a prior for η_{ik} to account for the variability in η_{ik} . However, one might be interested in both η_{ik} and ω_{ik} or a function of η_{ik} and ω_{ik} where ω_{ik} is often a function of the population size. Here we would need to model η_{ik} and ω_{ik} simultaneously. A general class of GLMM for η_{ik} and ω_{ik} could be

$$h_1(\eta_{ik}) = \mathbf{X}_1 \boldsymbol{\theta}_{1k} + \mathbf{S}_1 \mathbf{u}_{1k} + \mathbf{e}_{1k}, \quad (12)$$

$$h_2(\omega_{ik}) = \mathbf{X}_2 \boldsymbol{\theta}_{2k} + \mathbf{S}_2 \mathbf{u}_{2k} + \mathbf{e}_{2k}. \quad (13)$$

With $a = 1, 2$, $\mathbf{X}_a = \{x_{aik}\}$ and $\mathbf{S}_a = \{s_{aik}\}$ are known design matrices. The vector $\boldsymbol{\theta}_{ak}$ is the vector of fixed effects, \mathbf{u}_{ak} is the vector of random effects, and e_{aik} are independent residual effects and $e_{aik} \sim N(0, \delta_{ak}^{(e)})$. In addition, \mathbf{u}_{ak} and e_{aik} are assumed mutually independent.

3.2 Additional priors

To complete the Bayesian hierarchical model, we need to specify the priors for $(\boldsymbol{\theta}_{ak}, \mathbf{u}_{ak}, \delta_{ak}^{(e)})$, $a = 1, 2, k = 1, \dots, K$. The common prior for fixed effects $\boldsymbol{\theta}_{ak}$ is normal with a large variance or a constant prior. Random effects are often spatially correlated. The density may be of the CAR form whose joint density is given by equation (3) with $\mathbf{B}_{ak} = \mathbf{I} - \rho_{ak}\mathbf{C}$. Finally, a common prior for variance components $\delta_{ak}^{(e)}$ is an inverse gamma distribution.

To evaluate the posterior distribution, MCMC methods, such as Gibbs sampling, may be used to obtain samples from the posterior distribution. We give full conditional distributions in the general case below.

Fact 1 Let $(\Omega | \cdot)$ represent the conditional distributions of Ω given all other parameters and $[\Omega | \cdot]$ represent the conditional density. The conditional posterior densities of η_{ik} and ω_{ik} are as follows.

- i) $[\eta_{ik} | \cdot] \propto \exp(A_1(\phi_1)[y_{ik}\eta_{ik} - B_1(\eta_{ik}, n_{ik})] - 1/2\delta_{ik}^{(e)}[h_1(\eta_{ik}) - \mathbf{x}_{1i}'\boldsymbol{\theta}_{1k} - S_{1i}u_{1ik}]^2) h_1'(\eta_{ik})$ or equivalently $v_{1ik} = h_1(\eta_{ik})$ has the conditional density,

$$[v_{1ik} | \cdot] \propto \exp \left\{ \begin{aligned} &A_1(\phi_1)[y_{ik}h_1^{-1}(v_{1ik}) \\ &- B_1(h_1^{-1}(v_{1ik}), n_{ik})] \\ &- 1/2\delta_{ik}^{(e)}(v_{1ik} - \mathbf{x}_{1i}'\boldsymbol{\theta}_{1k} - S_{1i}u_{1ik})^2 \end{aligned} \right\}.$$

- ii) $[\omega_{ik} | \cdot] \propto \exp(A_2(\phi_2)[n_{ik}\omega_{ik} - B_2(\omega_{ik})] - 1/2\delta_{2k}^{(e)}[h_2(\omega_{ik}) - \mathbf{x}_{2i}'\boldsymbol{\theta}_{2k} - S_{2i}u_{2ik}]^2) h_2'(\omega_{ik})$ or equivalently $v_{2ik} = h_2(\omega_{ik})$ has the conditional density,

$$[v_{2ik} | \cdot] \propto \exp \left\{ \begin{aligned} &A_2(\phi_2)[n_{ik}h_2^{-1}(v_{2ik}) \\ &- B_2(h_2^{-1}(v_{2ik}))] \\ &- 1/2\delta_{2k}^{(e)}(v_{2ik} - \mathbf{x}_{2i}'\boldsymbol{\theta}_{2k} - S_{2i}u_{2ik})^2 \end{aligned} \right\}.$$

For $a = 1, 2$, we have the following conditional distributions.

- iii) $(\boldsymbol{\theta}_{ak} | \cdot)$ are mutually independent and have conditional posterior distributions

$$N_I \left(\begin{aligned} &\left(\frac{1}{\tau_{ak}} \mathbf{I} + \frac{1}{\delta_{ak}^{(e)}} \mathbf{X}_a' \mathbf{X}_a \right)^{-1} \\ &\left(\frac{1}{\delta_{ak}^{(e)}} \mathbf{X}_a' (\mathbf{v}_{ak} - \mathbf{u}_{ak}) + \frac{\mu_{ak}}{\tau_{ak}} \mathbf{1} \right), \\ &\left(\frac{1}{\tau_{ak}} \mathbf{I} + \frac{1}{\delta_{ak}^{(e)}} \mathbf{X}_a' \mathbf{X}_a \right)^{-1} \end{aligned} \right).$$

- iv) $(\mathbf{u}_{ak} | \cdot)$ are mutually independent and have conditional posterior distributions

$$N_I \left(\begin{aligned} &\left(\frac{1}{\delta_{ak}^{(e)}} \mathbf{S}_a' \mathbf{S}_a + \frac{1}{\delta_{ak}^{(u)}} \mathbf{B}_{ak} \right)^{-1} \\ &\left(\frac{1}{\delta_{ak}^{(e)}} \mathbf{S}_a' (\mathbf{v}_{ak} - \mathbf{X}_a \boldsymbol{\theta}_{ak}) \right), \\ &\left(\frac{1}{\delta_{ak}^{(e)}} \mathbf{S}_a' \mathbf{S}_a + \frac{1}{\delta_{ak}^{(u)}} \mathbf{B}_{ak} \right)^{-1} \end{aligned} \right).$$

- v) $(\delta_{ak}^{(e)} | \cdot)$ have posterior distributions

$$\text{Inverse Gamma} \left(\begin{aligned} &\left(\alpha_{ak}^{(e)} + I/2, \beta_{ak}^{(e)} \right) \\ &+ 1/2 (\mathbf{v}_{ak} - \mathbf{X}_a \boldsymbol{\theta}_{ak} - \mathbf{S}_a \mathbf{u}_{ak})' \\ &(\mathbf{v}_{ak} - \mathbf{X}_a \boldsymbol{\theta}_{ak} - \mathbf{S}_a \mathbf{u}_{ak}) \end{aligned} \right).$$

- vi) $(\delta_{ak}^{(u)} | \cdot) \sim \text{Inverse Gamma}(\alpha_{ak}^{(u)} + I/2, \beta_{ak}^{(u)} + 1/2 \mathbf{u}_{ak}' \mathbf{B}_{ak} \mathbf{u}_{ak})$.

- vii) $[\rho_{ak} | \cdot] \propto |\mathbf{B}_{ak}|^{1/2} \exp \left\{ -\frac{\mathbf{u}_{ak}' \mathbf{B}_{ak} \mathbf{u}_{ak}}{2\delta_{ak}^{(u)}} \right\}$.

- viii) If ϕ_1 has the prior density $p(\phi_1)$, then

$$[\phi_1 | \cdot] \propto p(\phi_1) \exp \left\{ \begin{aligned} &A_1(\phi_1)[y_{ik}h_1^{-1}(v_{1ik}) \\ &- B_1(h_1^{-1}(v_{1ik}), n_{ik})] \\ &+ C_1(v_{1ik}, n_{ik}, \phi_1) \end{aligned} \right\}.$$

- ix) If ϕ_2 has the prior density $p(\phi_2)$, then

$$[\phi_2 | \cdot] \propto p(\phi_2) \exp \left\{ \begin{aligned} &A_2(\phi_2)[n_{ik}h_2^{-1}(v_{2ik})] \\ &-B_2(h_2^{-1}(v_{2ik})) \\ &+C_2(n_{ik}, \phi_2) \end{aligned} \right\}.$$

Examining parts *i*), *ii*), and *vii*) of Fact 1, the conditional densities of η_{ik}, ω_{ik} and ρ_{ak} are often log-concave.

3.3 Estimating quantities of study domains

It is often of interest to estimate certain quantities related to the study domains. For instance, to estimate a quantity at domain *i*, let

$$\psi_i = f_i(\eta_{i1}, \omega_{i1}, \dots, \eta_{iK}, \omega_{iK}).$$

Bayesian estimates of ψ_i can be easily computed based on a random sample from the joint posterior. For example, let $(\eta_{ik}^{(l)}, \omega_{ik}^{(l)})$, $l = 1, \dots, L$ and $k = 1, \dots, K$ be the output from MCMC simulations, and define

$$\psi_i^{(l)} = f_i(\eta_{i1}^{(l)}, \omega_{i1}^{(l)}, \dots, \eta_{iK}^{(l)}, \omega_{iK}^{(l)}).$$

Given $\mathbf{y} = \{y_{ik}, i = 1, \dots, L, k = 1, \dots, K\}$, the posterior mean and variance of ψ_i , the general forms of (4) and (5), can be approximated by

$$\hat{E}(\psi_i | \mathbf{y}) = \frac{1}{L} \sum_{l=1}^L \psi_i^{(l)}$$

and

$$\hat{V}(\psi_i | \mathbf{y}) = \frac{1}{L-1} \sum_{l=1}^L \{\psi_i^{(l)}\}^2 - \frac{L}{L-1} \{\hat{E}(\psi_i | \mathbf{y})\}^2,$$

respectively.

3.4 Predicting quantities of study domains

Now let N_{ik} be the population size and Y_{ik}^* the response value of interest for those not in the sample from study domain *i* and design stratum *k*. We wish to predict the quantities $Y_i^* = \sum_{k=1}^K Y_{ik}^*$, the total response value in study domain *i*. For given n_{ik}, Y_{ik}^* should be of the same family as Y_{ik} such that

$$g_3(y_{ik}^* | \eta_{ik}, n_{ik}, N_{ik}, \phi_1) = \exp \left[\begin{aligned} &A_1(\phi_1)\{y_{ik}^* \eta_{ik}\} \\ &-B_1(\eta_{ik}, N_{ik} - n_{ik}) \\ &+C_1(y_{ik}^*, n_{ik} - n_{ik}, \phi_1) \end{aligned} \right].$$

This is simply (9) with n_{ik} replaced by $N_{ik} - n_{ik}$. To simplify notation, let ξ denote the parameters in the model, \mathbf{d} represent the data. (In our case here $\mathbf{d} = \{(y_{ik}, n_{ik}): i = 1, \dots, L, k = 1, \dots, K\}$ and ξ might include the parameters in modeling both y_{ik} and n_{ik} .) Under the prior $\pi(\xi)$ (either proper or improper), we have the posterior density $[\xi | \mathbf{d}] \propto f(\mathbf{d} | \xi) \pi(\xi)$.

Assume that a further observation y_{ik}^* follows $g_3(y_{ik}^* | \mathbf{d}, \xi)$ that may be dependent on \mathbf{d} . The predictive density of y_{ik}^* given \mathbf{d} is written as

$$[y_{ik}^* | \mathbf{d}] = \int_{\xi} g_3(y_{ik}^* | \xi, \mathbf{d}) [\xi | \mathbf{d}] d\xi.$$

Under the squared error loss function, the best predictor is then the predictive mean given by

$$E(y_{ik}^* | \mathbf{d}) = \int_{\xi} \left\{ \int_{y_{ik}} y_{ik}^* g_3(y_{ik}^* | \xi, \mathbf{d}) dy_{ik}^* \right\} [\xi | \mathbf{d}] d\xi.$$

Similarly, the best predictor of $h(y_{i1}^*, \dots, y_{iK}^*)$ given \mathbf{d} is then

$$E\{h(y_{i1}^*, \dots, y_{iK}^*) | \mathbf{d}\} = \int_{\xi} E\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi, \mathbf{d}\} [\xi | \mathbf{d}] d\xi, \quad (14)$$

where

$$E\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi, \mathbf{d}\} = \int_{(y_{i1}^*, \dots, y_{iK}^*)} h(y_{i1}^*, \dots, y_{iK}^*) \prod_{k=1}^K [y_{ik}^* | \xi, \mathbf{d}] dy_{i1}^* \dots dy_{iK}^*. \quad (15)$$

For the distribution family $g_3(y_{ik}^* | \xi, \mathbf{d})$, the right hand side of (15) often has a closed form expression.

Bayesian predictions of (14) can be easily computed based on a random sample from the joint posterior of ξ . For example, let $\xi^{(l)}$, $l = 1, \dots, L$, be the output from MCMC simulations. The posterior predictive mean (14) can then be approximated by

$$\hat{E}\{h(y_{i1}^*, \dots, y_{iK}^*) | \mathbf{d}\} = \frac{1}{L} \sum_{l=1}^L E\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi^{(l)}, \mathbf{d}\}.$$

The posterior predictive variance of $h(y_{i1}^*, \dots, y_{iK}^*)$ given \mathbf{d} may be written as

$$V\{h(y_{i1}^*, \dots, y_{iK}^*) | \mathbf{d}\} = E[V\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi, \mathbf{d}\}] + V[E\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi, \mathbf{d}\}].$$

This predictive variance can be approximated by

$$\begin{aligned} \hat{V}\{h(y_{i1}^*, \dots, y_{iK}^*) | \mathbf{d}\} &= \frac{1}{L} \sum_{l=1}^L V\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi^{(l)}, \mathbf{d}\} \\ &+ \frac{1}{L} \sum_{l=1}^L [E\{h(y_{i1}^*, \dots, y_{iK}^*) | \xi^{(l)}, \mathbf{d}\} - \hat{E}\{h(y_{i1}^*, \dots, y_{iK}^*) | \mathbf{d}\}]^2. \end{aligned}$$

4. Comments

In this article, we developed a bivariate GLMM containing two unknown canonical parameters. This was necessary to obtain estimates when the sample size was random and estimates N_{ik} were required in addition to estimates of η_{ik} . The model was built using two simultaneous GLMMs in a Bayesian hierarchical structure.

The proposed model has the advantage of being applicable to a wide array of problems. Introducing a random sample size and estimating the population sizes are useful techniques for many applications.

Naturally, we think that there is an inverse relationship between hunting success rates and the number of trips taken to a county. In modeling this, we may think of a two-fold CAR model to view the relationship between η_{ik} and N_{ik} such as that of Kim, Sun and Tsutakawa (2001). A multivariate CAR model may be another approach to address this situation.

For each of the spatial models, we have assumed a common correlation, ρ_k , across the entire state. It may be more reasonable to include additional correlation terms in different productivity regions defined by the Missouri Department of Conservation. This would be an interesting and complicated addition to the model. In addition, the spatial structure used in this paper is similar to that of He and Sun (2000) and Oleson and He (2004) where this spatial modeling worked well.

It may be useful to include the distance from hunter's home to the hunting location to help estimate hunting pressure. Most hunters stay close to home when hunting and this information could be incorporated into the hierarchical framework.

Note that the estimated harvest is higher than the check station harvest. This is partly because more successful hunters tend to reply to the mail survey. We are conducting research adjusting for the nonresponse bias.

Acknowledgements

The work was partially supported by Federal Aid in Wildlife Restoration Project W-13-R. Sun's research was supported by the National Science Foundation grant SES-0351523, and NIH grant R01-CA100760. The authors would like to thank Dr. Larry Vangilder and Mr. Jeff Beringer of the Missouri Department of Conservation for their helpful comments. The authors wish to express deep appreciation to the editors, an associate editor, and two referees for constructive comments on earlier versions of the paper.

References

- Agresti, A. (2002). *Categorical data analysis*. New York: John Wiley & Sons, Inc.
- Basag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Methodological*, Series B, 36, 192-236.
- Carlin, B.P., and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall Ltd.
- Clayton, D., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc. Third Edition.
- Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.
- Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New developments in survey sampling*, (Eds., N.L. Johnson and H.J. Smith). New York: Wiley-Interscience, 629-651.
- Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, London, U.K.
- Ghosh, M., Natarajan, K., Stroud, T. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Gilks, W.R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.
- He, C.Z., and Sun, D. (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics*, 5, 223-236.
- He, C.Z., and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.
- Heidelberger, P., and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.
- Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.
- Jiang, J., Lahiri, P. and Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M -estimation. *The Annals of Statistics*, 30, 1782-1810.
- Kim, H., Sun, D. and Tsutakawa, R.K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96, 1506-1521.
- Kimmel, R.O. (2001). Regulating spring wild turkey hunting based on population and hunting quality. In *Proceeding of the National Wildlife Turkey Symposium*, 8, 243-250.
- Lohr, S.L. (1999). *Sampling: Design and analysis*. Duxbury Press.
- Malec, D., Sedransk, J., Moriarity, C.L. and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

- Oleson, J.J., and He, C.Z. (2004). Space-time modeling for the Missouri Turkey Hunting Survey. *Environmental and Ecological Statistics*, 11, 85-101.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley-Interscience.
- Rao, J.N.K. (2005). Inferential issues in small area estimation: Some new developments. *Statistics in Transition*, 7, 513-526.
- Robert, C.P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.
- Smith, B.J. (2005). Bayesian Output Analysis program (BOA), version 1.1.5. <http://www.public-health.uiowa.edu/boa>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (Pkg: 583-639). *Journal of the Royal Statistical Society, Methodological*, Series B, 64, 583-616.
- Sun, D., Speckman, P.L. and Tsutakawa, R.K. (2000). Random effects in generalized linear mixed models (GLMMs). In *Generalized Linear Models: A Bayesian Perspective*, (Eds., D.K. Dey, S.K. Ghosh and B.K. Mallick). New York : Marcel Dekker, 23-39.
- Vangilder, L.D., Sheriff, S.L. and Olsen, G.S. (1990). Characteristics, attitudes, and preferences of Missouri's spring turkey hunters. In *Proceeding of the National Wildlife Turkey Symposium*, 6, 167-176.
- Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Statistical Methodology*, Series B, 62, 159-180.
- Woodard, R., He, C.Z. and Sun, D. (2003). Bayesian estimation of hunting success rate and harvest for spatially correlated post-stratified data. *Biometrical Journal*, 45, 985-1005.
- You, Y., and Rao, J. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111, 197-208.

Small area estimation of average household income based on unit level models for panel data

Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacei¹

Abstract

The European Community Household Panel (ECHP) is a panel survey covering a wide range of topics regarding economic, social and living conditions. In particular, it makes it possible to calculate disposable equivalized household income, which is a key variable in the study of economic inequity and poverty. To obtain reliable estimates of the average of this variable for regions within countries it is necessary to have recourse to small area estimation methods. In this paper, we focus on empirical best linear predictors of the average equivalized income based on "unit level models" borrowing strength across both areas and times. Using a simulation study based on ECHP data, we compare the suggested estimators with cross-sectional model-based and design-based estimators. In the case of these empirical predictors, we also compare three different MSE estimators. Results show that those estimators connected to models that take units' autocorrelation into account lead to a significant gain in efficiency, even when there are no covariates available whose population mean is known.

Key Words: European Community Household Panel; Average equivalized income; Linear mixed models; Empirical best linear unbiased predictor; MSE estimation.

1. Introduction

In recent years, the academic world has taken an increasing interest in the analysis of regional economic disparities that represent a serious challenge to the promotion of national economic growth, and thus to social cohesion. This is particularly true within the European Union, where regional disparities are a distinguishing feature of the economic landscape. This renewed interest in local economies has produced a growing demand for regional statistical information and has stimulated research on income distribution, poverty and social exclusion at the sub-national level.

In the 1990s, Eurostat (the EU's Statistics Bureau) launched the European Community Household Panel (ECHP), an annual panel survey of European households conducted using standardised methods throughout the EU's various member countries (Betti and Verma 2002; Eurostat 2002). The ECHP terminated in 2001, after eight waves. Currently, it is being replaced by the Survey on Income and Living Conditions in the Community (EU-SILC), which resembles the ECHP in many ways, but for which no data has yet been published. The ECHP panel survey covered a wide range of topics and, in particular, it made it possible to calculate disposable equivalized household income, which constitutes a key variable in the study of economic equity and poverty.

The ECHP was designed to provide reliable estimates for large areas within countries called NUTS1 (NUTS stands for the "Nomenclature of Territorial Units for Statistics")

which is defined according to certain principles described on the EUROSTAT web site http://europa.eu.int/comm/eurostat/ramon/nuts/home_regions_en.html). Unfortunately NUTS1 correspond to areas (five groups of Administrative Regions in the Italian case) that are too large to effectively measure local area income disparity or to provide useful information for the purposes of regional governance. Therefore, to obtain estimates for a finer geographic detail, a small area estimation method has to be used and the problem is to select an appropriate and effective method.

In this paper, in order to combine information from past surveys, related auxiliary variables and small areas, we consider several possible extensions of the well-known unit level nested error regression model (see Battese, Harter and Fuller 1988) for the estimation of the average of household equivalized income. Using ECHP panel survey data, we illustrate how such model could be potentially useful in improving the efficiency of small area estimates by exploiting the correlation of individual household incomes over time.

In section 2, we present a general set-up for small area estimation using panel survey data and briefly review both design-based and model-based small area estimation methods. In this section, we develop empirical best linear unbiased predictors (EBLUP) and their mean squared error (MSE) estimators for selected unit level cross-sectional and time series models using the available theory on EBLUP for small area estimation (see Rao 2003, and Jiang and Lahiri 2006a, for details). We note that cross-sectional and time series models were considered in the small area literature,

1. Enrico Fabrizi, DMSIA, University of Bergamo, via dei Caniana 2, 24127, Bergamo, Italy. E-mail: enrico.fabrizi@unibg.it; M.R. Ferrante, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy. E-mail: ferrante@stat.unibo.it; S. Pacei, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy. E-mail: pacei@stat.unibo.it.

but only in the context of area level modelling (see Rao and Yu 1994; Ghosh, Nangia and Kim 1996; Datta, Lahiri, Maiti and Lu 1999; Datta, Lahiri and Maiti 2002; Pfeffermann 2002; among others).

In section 3, we briefly review ECHP survey and describe how we use this survey data to conduct a Monte Carlo simulation study to compare different small area estimators and their MSE estimators. In section 4, we report results from the Monte Carlo simulation experiment. We note that the simulation experiment is aimed at evaluating design-based properties of all estimators, even if they are derived as model based predictors. We observed that the EBLUPs perform very well compared to the design-based estimators even though our pseudo-population exhibits signs of non-normality. The non-normality of the pseudo-population, however, seems to affect the efficiency of the MSE estimators. In our simulation, the Taylor series (see Prasad and Rao 1990; Datta and Lahiri 1999, among others) and the parametric bootstrap (see Butar and Lahiri 2003) MSE estimators are found out to be more sensitive to the non-normality than the jackknife method of Jiang, Lahiri and Wan (2002). We end the paper with a few concluding remarks.

2. The small area estimation methods considered

To describe sample data, let y_{diti} denote the value of a study variable for the i^{th} unit belonging to the d^{th} small area for time t ($d = 1, \dots, m$; $t = 1, \dots, T$; $i = 1, \dots, n_d$). Moreover, let \mathbf{x}_{diti} be the vector of covariates' values associated with each y_{diti} (and whose first element is equal to 1), and let $\mathbf{X}_s = \{\mathbf{x}_{diti}\}$ be the $n \times p$ matrix of covariates' values for the whole sample ($n = \sum_{d,t} n_{dt}$). Let us suppose that we are interested in predicting small area means for the target variable at final time T : \bar{Y}_{dT} , ($d = 1, \dots, m$). Let us also suppose that the vectors of mean population values of covariates are known for time T ; we denote these vectors by $\bar{\mathbf{X}}'_{dT}$ ($d = 1, \dots, m$).

2.1 Design-based estimators

A first solution to the small area estimation problem is to use direct estimators, that is, estimators employing only y values obtained from the area (and time) which the parameter refers to. The simplest of direct estimators of the population mean is the weighted mean. We denote this direct estimator as $\bar{y}_{dT, \text{DIR}}$ ($d = 1, \dots, m$) and we will be using it as a benchmark in the following sections.

Synthetic estimators may be generally defined as unbiased estimators for a large area with acceptable standard errors. They are used to calculate estimates for small areas, under the hypothesis that small areas have the

same characteristics as larger ones. Moreover, when information about auxiliary variables is available, a particular synthetic estimator, the regression estimator, may be obtained by fitting a regression model to all sample data. Note that the synthetic estimator is area specific with respect to the auxiliary variables but not with respect to the study variable.

For instance, if we consider only those observations from the last wave ($t = T$), the simple regression model would be given by:

$$y_{dT} = \mathbf{x}'_{dT} \beta + e_{dT} \\ E(e_{dT}) = 0, \quad E(e_{dT}^2) = \pi^2.$$

To take account of the complexity of the sampling design, the weighted least squares estimate $\hat{\beta}_w$ of β may be obtained, and thus the synthetic regression estimator will be given by:

$$\bar{y}_{dT, \text{RSYN}} = \bar{\mathbf{X}}'_{dT} \hat{\beta}_w, \quad d = 1, \dots, m. \quad (1)$$

Synthetic estimators usually display very low variances, but they may be severely biased whenever the model holding for the whole sample does not properly fit area-specific data. Composite estimators are weighted averages of a direct and a synthetic estimator. We consider the composite estimator:

$$\bar{y}_{dT, \text{COMP}} = \phi_{dT} \bar{y}_{dT, \text{DIR}} + (1 - \phi_{dT}) \bar{y}_{dT, \text{RSYN}}, \quad (2)$$

where

$$\phi_{dT} = \frac{\text{MSE}_D(\bar{y}_{dT, \text{RSYN}})}{\text{MSE}_D(\bar{y}_{dT, \text{DIR}}) + \text{MSE}_D(\bar{y}_{dT, \text{RSYN}})}$$

and MSE_D signifies that the mean square error is evaluated in relation to the randomization distribution. This choice of ϕ_{dT} leads to composite estimators $\bar{y}_{dT, \text{COMP}}$ that are approximately optimal in terms of MSE_D (see Rao 2003, section 4.3). In practice, the quantities in the formula for ϕ_{dT} 's are unknown and may be estimated from the data. Unbiased and consistent estimators can be obtained for $\text{MSE}_D(\bar{y}_{dT, \text{DIR}}) = V_D(\bar{y}_{dT, \text{DIR}})$ using standard formulas. An approximately design unbiased estimator of $\text{MSE}_D(\bar{y}_{dT, \text{RSYN}})$ can be obtained using the formulas discussed in Rao (2003, section 4.2.4). In particular, we calculate the approximation:

$$\text{mse}_D(\bar{y}_{dT, \text{RSYN}}) \approx (\bar{y}_{dT, \text{RSYN}} - \bar{y}_{dT, \text{DIR}})^2 - v_D(\bar{y}_{dT, \text{DIR}}),$$

where mse_D and v_D stand for the estimators of the corresponding MSE_D and V_D . In particular, v_D is the ordinary design unbiased estimator of V_D . We then take its average over d , as usual, in order to obtain a more stable estimator. In fact, one problem with mse_D is that it can even be negative.

Moreover, a modified direct estimator borrowing strength over areas for estimating the regression coefficient can be used to improve estimator reliability. If auxiliary information is available, the generalized regression estimator (GREG),

$$\bar{y}_{dT, \text{GREG}} = \bar{\mathbf{x}}'_{dT} \hat{\beta}_w + \frac{\sum_{j \in s_{dT}} w_j e_j}{\sum_{j \in s_{dT}} w_j}, \quad (3)$$

approximately corrects the bias of the synthetic estimator by means of the term $(\sum_{j \in s_{dT}} w_j)^{-1} \sum_{j \in s_{dT}} w_j e_j$, based on regression residuals e_j .

2.2 Model-based estimators

The model-based estimators we have considered are based on the specification of explicit models for sample data which approximate a hypothetical data-generating process. As a consequence, the problem of estimating \bar{y}_{dT} comes down to one of prediction. Moreover, mean square errors and other statistical properties of estimators are usually evaluated with respect to the data-generating process. We have focused here on "unit level" models based on models relating y_{diti} to a vector of covariates \mathbf{x}_{diti} . The use of explicit models has several advantages, the most important of which being the opportunity to test underlying assumptions.

In the estimation of the small area means or totals of continuous variables, linear mixed models are very often used. A general linear mixed model can be described as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1 \mathbf{v}_1 + \dots + \mathbf{Z}_s \mathbf{v}_s + \mathbf{e}, \quad (4)$$

where $\mathbf{y} = \{y_{diti}\}$ is the n -vector of sample observations, β a $p \times 1$ vector of fixed effects, \mathbf{v}_j is a $q_j \times 1$ vector of random effects ($j = 1, \dots, s$), $\mathbf{e} = \{e_{diti}\}$ a vector of errors; \mathbf{X} is assumed of rank p , $\mathbf{Z}_i = \{\mathbf{z}'_{diti}\}$ is a $n \times q_j$ matrix of incidence of the j^{th} random effect. We assume that $E(\mathbf{v}_j) = 0$, $V(\mathbf{v}_j) = \mathbf{G}_j$, $E(\mathbf{e}) = 0$, $V(\mathbf{e}) = \mathbf{R}$ (all expectations are wrt. model (4)) and that $\mathbf{v}_1, \dots, \mathbf{v}_s, \mathbf{e}$ are mutually independent.

As a consequence, the variance-covariance matrix of \mathbf{y} is given by:

$$\mathbf{V} = V(\mathbf{y}) = \sum_{j=1}^s \mathbf{Z}_j \mathbf{G}_j \mathbf{Z}'_j + \mathbf{R} = \mathbf{ZGZ}' + \mathbf{R},$$

where $\mathbf{Z} = [\mathbf{Z}_1 | \dots | \mathbf{Z}_s]$. It is usually assumed that matrices \mathbf{G} , \mathbf{R} depend on a k -vector of variance components ψ , and so we can write $\mathbf{V}(\psi) = \mathbf{ZG}(\psi)\mathbf{Z}' + \mathbf{R}(\psi)$.

Note that at the level of individual observations, the model (4) can be rewritten as $y_{diti} = \mathbf{x}'_{diti} \beta + \mathbf{z}'_{1diti} \mathbf{v}_1 + \dots + \mathbf{z}'_{sditi} \mathbf{v}_s + e_{diti}$.

We consider different specifications for linear mixed models, all of which can be viewed as special cases of the general model (4). For the sake of simplicity, we have adopted a unit level notation when describing the models considered. The first model:

$$MM1: y_{diti} = \mathbf{x}'_{diti} \beta + v_d + \alpha_t + e_{diti}, \quad (5)$$

may be obtained from formula (4) setting $s = 2$, $q_1 = m$, $q_2 = T$, $\mathbf{G}_1 = \sigma_v^2 \mathbf{I}_m$, $\mathbf{G}_2 = \sigma_\alpha^2 \mathbf{I}_T$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. It includes independent area and time effects, and therefore area effects are assumed not to evolve over time. This random effects structure corresponds to the assumption of a constant covariance between units that belong to the same area, observed at two different points in time.

The second model:

$$MM2: y_{diti} = \mathbf{x}'_{diti} \beta + \delta_{dt} + e_{diti}, \quad (6)$$

corresponds to the particular case in which $s = 1$, $q_1 = mq$, $\mathbf{G}_1 = \sigma_\delta^2 \mathbf{I}_{mq}$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. The effects of interaction between area and time are introduced, that is, we assume there are area effects which are not constant over time.

The third model:

$$MM3: y_{diti} = \mathbf{x}'_{diti} \beta + v_d + \alpha_t^* + e_{diti}, \quad (7)$$

is obtained setting $s = 2$, $q_1 = m$, $q_2 = T$, $\mathbf{G}_1 = \sigma_v^2 \mathbf{I}_m$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$, while the generic element $g_2(h, k)$ of \mathbf{G}_2 is $g_2(h, k) = \sigma_\alpha^2 \rho_\alpha^{[h-k]}$, $h, k = 1, \dots, T$. There are independent area and time effects, just as in *MM1*, but the time effects are assumed to follow an AR(1) process.

The fourth model:

$$MM4: y_{diti} = \mathbf{x}'_{diti} \beta + \delta_{dt}^* + e_{diti}, \quad (8)$$

is similar to model *MM2* in that it is characterized by time varying area effects, but the further assumption that such effects follow an AR(1) process is also introduced. Thus, provided we order observations by area, with respect to the general formula (4) we have $s = 1$, $q_1 = mq$, $\mathbf{G}_1 = \text{diag}(\mathbf{G}_{1d})$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$ where \mathbf{G}_{1d} , $d = 1, \dots, m$, is a $T \times T$ matrix the generic element $g_{1d}(h, k) = \sigma_\delta^2 \rho_\delta^{[h-k]}$, $h, k = 1, \dots, T$.

The last specification:

$$MM5: y_{diti} = \mathbf{x}'_{diti} \beta + v_d + \alpha_t + e_{diti}^*, \quad (9)$$

may be obtained by (4) setting $s = 2$, $q_1 = m$, $q_2 = T$, $\mathbf{G}_1 = \sigma_v^2 \mathbf{I}_m$, $\mathbf{G}_2 = \sigma_\alpha^2 \mathbf{I}_T$. Provided we order observations by household and time, $\mathbf{R} = \text{diag}(\mathbf{R}_{dt})$ where \mathbf{R}_{dt} is a $T \times T$ matrix whose generic element is given by $r_{dt}(h, k) = \sigma_e^2 \rho_e^{[h-k]}$, $h, k = 1, \dots, T$. There are independent area and time effects like in *MM1*, but errors are assumed to be autocorrelated according to an AR(1) process.

In order to evaluate the impact that using past survey waves has on the efficiency of estimator, a cross-sectional linear mixed model (*SMM*) using data from the last wave T only, has been taken as the benchmark:

$$SMM: y_{dTi} = \mathbf{x}'_{dTi} \beta + \vartheta_d + e_{dTi} \quad (10)$$

with $\vartheta_d \sim N(0, \sigma_\vartheta^2)$, $e_{dTi} \sim N(0, \sigma_e^2)$.

This is also a particular case of (4) obtained for $s = 2$, $q_1 = m$, $\mathbf{G}_1 = \sigma_\vartheta^2 \mathbf{I}_m$ and $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Note that (10) is the standard nested error regression model of Battese *et al.* (1988).

We also consider the corresponding random error variance linear models (see Rao 2003; section 5.5.2) obtained by replacing $\mathbf{x}'_{dTi} \beta$ in formulas (5) - (10) with a general intercept θ . These models will be denoted as *MM1**, *MM2**, *MM3**, *MM4**, *MM5**, *SMM**. All the assumptions made regarding random effects and residuals remain unchanged. This latter group of models enables us to explore the gains in efficiency obtained by exploiting the repetition of the observation on the same unit when no covariates are available at the population level.

In small area estimation, the aim is to predict scalar linear combinations of fixed and random effects of the type $\eta = \mathbf{m}'\beta + \mathbf{k}'\mathbf{v}$ where \mathbf{m} and \mathbf{k} are $p \times 1$ and $q \times 1$ vectors respectively, with $q = \sum_j q_j$. The best linear unbiased predictor (BLUP) of η can be obtained by estimating (β, \mathbf{v}) minimizing the model MSE among all linear estimators:

$$\tilde{\eta}^{BLUP}(\psi) = \mathbf{m}'\tilde{\beta}(\psi) + \mathbf{k}'\tilde{\mathbf{v}}(\psi). \quad (11)$$

When the variance components in ψ are unknown, they may be estimated from the data and substituted into formula (11), thus obtaining "empirical BLUP" $\tilde{\eta}^{EBLUP}(\hat{\psi}) = \mathbf{m}'\hat{\beta}(\hat{\psi}) + \mathbf{k}'\hat{\mathbf{v}}(\hat{\psi})$ (see Rao 2003, chapter 6, and Jiang and Lahiri 2006b for details).

As far as the estimation of ψ is concerned, a number of methods have been proposed in the literature, such as Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) which assume the normality of random terms, and the MINQUE proposed by Rao (1972) which is non-parametric. In the present work we have opted for the REML method, thus assuming normality.

2.3. Measures of uncertainty associated with predictors based on linear mixed models

The difficult problem of estimating the MSE of EBLUP estimators, taking the variability of the estimated variance and covariance components into account, has been faced in the small area literature by adopting diverse approaches.

One popular method is based on the Taylor series approximation of MSE under normality (Prasad and Rao

1990; Datta and Lahiri 1999). More recently, due to the advent of high-speed computers and powerful software, resampling methods have been proposed. For instance, Butar and Lahiri (2003) introduce a parametric bootstrap method based on the assumption of normality, but analytically less onerous than the Taylor series method. Jiang *et al.* (2002) discuss a general jackknife method, which requires a distributional assumption weaker than normality (posterior linearity). We aim to empirically compare the performance of these three estimators within a context where the number of areas is moderate and the assumption of normality may not hold perfectly true. The following is a short description of the three estimation approaches.

Let us define $MSE[\tilde{\eta}^{EBLUP}(\hat{\psi})] = E(\tilde{\eta}^{EBLUP}(\hat{\psi}) - \eta)^2$, where expectation refers to model (4). It is possible to show that, under normality,

$$\begin{aligned} MSE[\tilde{\eta}^{EBLUP}(\hat{\psi})] &= g_1(\psi) + g_2(\psi) + E(\tilde{\eta}^{EBLUP} - \tilde{\eta}^{BLUP})^2 \end{aligned} \quad (12)$$

where $g_1(\psi) = \mathbf{k}'(\mathbf{G} - \mathbf{GZ}'\mathbf{V}^{-1}\mathbf{ZG})\mathbf{k}'$ and $g_2(\psi) = \mathbf{d}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}$, with $\mathbf{d} = \mathbf{m}' - \mathbf{k}'\mathbf{GZ}'\mathbf{V}^{-1}\mathbf{X}$ (see Rao 2003, chapter 3). Using the following approximation, based on a Taylor series argument

$$\begin{aligned} E(\tilde{\eta}^{EBLUP} - \tilde{\eta}^{BLUP})^2 &\approx \text{tr}[(\partial \mathbf{b}' / \partial \psi) \mathbf{V} (\partial \mathbf{b}' / \partial \psi)' \bar{\mathbf{V}}(\psi)] = g_3(\psi) \end{aligned}$$

where $\mathbf{b}' = \mathbf{k}'\mathbf{GZ}'\mathbf{V}^{-1}$, a second order approximation to (12) can be found:

$$MSE[\tilde{\eta}^{EBLUP}(\hat{\psi})] \approx g_1(\psi) + g_2(\psi) + g_3(\psi). \quad (13)$$

Note that here \approx means that the omitted terms are of order $o(m^{-1})$. An asymptotically unbiased estimator of (13), based on Prasad and Rao (1990), is given by

$$mse_{PR}(\tilde{\eta}^{EBLUP}) = g_1(\hat{\psi}) + g_2(\hat{\psi}) + 2g_3(\hat{\psi}). \quad (14)$$

Datta and Lahiri (1999) show that, under normality and REML or ML estimation of ψ , $mse_{PR}(\tilde{\eta}^{EBLUP})$ estimates $MSE[\tilde{\eta}^{EBLUP}(\hat{\psi})]$ with a bias of order $o(m^{-1})$.

Butar and Lahiri (2003) propose a parametric bootstrap estimation of (13) under the assumption of normality. We adapt their estimator to the models we are analysing, assuming the following bootstrap model:

- i) $\mathbf{y}^* | \mathbf{v}^* \sim N[\mathbf{X}\hat{\beta} + \mathbf{Z}\mathbf{v}^*, \mathbf{R}(\hat{\psi})]$
- ii) $\mathbf{v}^* \sim N[\mathbf{0}, \mathbf{G}(\hat{\psi})]$

where $\mathbf{v} = (v_1, \dots, v_s)'$. The parametric bootstrap is then used twice, once to estimate the first two terms of (13), thus

correcting the bias of $g_1(\hat{\psi}) + g_2(\hat{\psi})$, and once to estimate $g_3(\psi)$.

The following estimator of (13) is proposed:

$$\begin{aligned} \text{mse}_{\text{BL}}(\tilde{\eta}^{\text{EBLUP}}) \\ = 2[g_1(\hat{\psi}) + g_2(\hat{\psi}) - E_B[g_1(\hat{\psi}^*) + g_2(\hat{\psi}^*)] \\ + E_B[\tilde{\eta}(\mathbf{y}, \hat{\beta}(\hat{\psi}^*), \hat{\psi}^*) - \tilde{\eta}(\mathbf{y}, \hat{\beta}(\hat{\psi}), \hat{\psi})] \end{aligned} \quad (16)$$

where $\hat{\psi}^*$ is the same as $\hat{\psi}$ except that it is calculated on \mathbf{y}^* instead of \mathbf{y} , and E_B is the expected value with regard to the bootstrap model (15).

The bootstrap estimator (16) does not require the analytical derivation of $g_3(\hat{\psi})$ which can be rather laborious when \mathbf{G} and \mathbf{R} have complicated structures.

Jiang *et al.* (2002) introduced a general jackknife estimator for the variance of empirical best predictors in linear and non-linear mixed models with M -estimation. In the problem we are investigating here, the estimator they propose can be written as:

$$\begin{aligned} \text{mse}_{\text{JLW}}(\tilde{\eta}^{\text{EBLUP}}) = g_1(\hat{\psi}) - \frac{m-1}{m} \sum_{j=1}^m [g_1(\hat{\psi}_{-j}) - g_1(\hat{\psi})] \\ + \frac{m-1}{m} \sum_{j=1}^m (\tilde{\eta}_{-j}^{\text{EBLUP}} - \tilde{\eta}^{\text{EBLUP}})^2 \end{aligned} \quad (17)$$

where $\hat{\psi}_{-j}$ is the estimate of ψ calculated by using all data except those from the j^{th} area. Similarly, $\tilde{\eta}_{-j}^{\text{EBLUP}} = \tilde{\eta}^{\text{EBLUP}}(\mathbf{y}_{-j}, \hat{\beta}(\hat{\psi}_{-j}), \hat{\psi}_{-j})$.

It is worth pointing out that, on the basis of the simulation results reported in Jiang *et al.* (2002), mse_{JLW} is deemed to be more robust than mse_{PR} with regard to departures from the assumption of normality, which can also be expected to be crucial for mse_{BL} .

3. The simulation study based on the European Household Community Panel data

The target population of the ECHP survey consists of all resident households of a large subset of the EU member countries. Although general survey guidelines were issued by Eurostat, a certain degree of flexibility was allowed, so there are some differences in the sampling design across countries. As far as Italy is concerned, the survey is based on a stratified two stage design, in which strata were formed by grouping the PSUs (municipalities) according to geographic region (NUTS2) and demographic size. For more details of the survey, see Eurostat (2002).

The ECHP deals with unit non-response, sample attrition and new entries using weighting and imputation. As attrition could lead to biased estimates of income if it does not appear at random, the effect of poverty on dropout propensity has been investigated (Rendtel, Behr and Sisto

2003; Vandecasteele and Debels 2004), and the results of these studies show that in the case of some countries, including Italy, this effect disappears under the control of weighting variables.

We have focused our attention on the eight ECHP waves available for Italy (1994-2001). Given that our aim is to assess whether the use of several successive observations of the same household could be profitable for the purposes of small area estimation, we have overlooked the problem of attrition and only considered those households that participate to the survey for all waves.

Our target variable is disposable, post-tax household income at the time of the last wave (2001). In studies of poverty and inequality, income is often equalized according to an equivalence scale in order to avoid comparison problems caused by differences in the composition of households. We consider the widely-used modified OECD scale, also adopted by Eurostat (2002) in its publications on income, poverty and social exclusion. According to this scale, equalized income is calculated by dividing disposable household income by the number (k) of "equivalent adults", defined as $k = 1 + 0.5a + 0.3c$, where a is the number of adults other than the "head of the household" and c is the number of children aged 13 or less. In general, the equalized income can be perceived as the amount of income that an individual, living alone, should dispose of in order to attain the same level of economic wellbeing he/she enjoys in his/her household.

Of the many covariates available in the bountiful ECHP questionnaire, we have chosen only those for which area means were available from the 2001 Italian Census results. Thus the chosen covariates are: the percentage of adults; the percentage of employed; the percentage of unemployed; the percentage of people with a high/medium/low level of education in the household; household typology (presence of children, presence of aged people, *etc.*); the number of rooms per-capita and the tenure status of the accommodation (rented, owned *etc.*).

As we have said, the aim of this paper is to compare the performance of different estimators in the controlled environment of a simulation exercise. A number of works in the literature have compared small area estimators using Monte Carlo experiments in which samples are drawn from synthetic populations based either on Censuses (Falorsi, Falorsi and Russo 1994; Ghosh *et al.* 1996) or on the replication of sample units' records (Falorsi, Falorsi and Russo 1999; Lehtonen, Särndal and Veijanen 2003; Singh, Mantel and Thomas 1994). Since household income is not measured by the Italian Census (nor is it given by the results of other Censuses conducted by EU countries), we treated the ECHP survey data as the pseudo-population and then draw samples using stratified probability proportional to

size sampling, the size variable being given by survey weights. This solution may not be as good as that of using data from a real Census population, but it is hopefully more realistic than generating population values of household income from a parametric model.

Monte Carlo samples of 1,000 (roughly 15% of the actual ECHP sample size) were drawn from the synthetic population by stratified random sampling without replacement, with strata given by the 21 NUTS2 regions. Thus these regions are treated as planned domains (as in the ECHP) for which sample size in the small areas is established beforehand, so that the sampling fractions reflect the over-sampling of smaller regions exactly as they do in the actual ECHP sampling design. The region-specific sample sizes we obtained range from 14 to 112, being on average equal to 48. Therefore in our simulation $n = 1,000$, the number of small areas corresponds to that of the Italian Regions ($m = 21$) and the number of points in time corresponds to the ECHP available waves ($T = 8$).

The distribution of equivalized household income in our pseudo-population (that is the distribution obtained by weighted estimation from the ECHP sample data) is characterized by an overall mean of 22,547 Euros and a coefficient of variation of 0.59. The distribution is positively skewed (even though skewness is not extreme: skewness coefficient $\gamma_1 = \mu_3 / \sigma^3 \cong 2.5$) and kurtosis ($\kappa = \mu_4 / \sigma^4 \cong 14.3$). The difference between mean and median is 9% of the mean. An interesting feature is given by the large disparities among administrative regions (that are the small areas of interest in our study). The mean of the equivalized household income ranges from 16,604 to 27,011, that is the most affluent area has a mean equivalent income 62% higher than the poorest one. Also the coefficient of variation (ranging from 0.28 to 0.84), skewness (γ_1 ranging from 0.1 to 4.6) and kurtosis (κ ranging from -0.7 to 32.9) show that the distribution of our target variable is quite a bit different in different areas.

To motivate the selected specifications of the random effects part of the considered linear mixed models (see section 2.2), an approach often recommended in textbooks (see Verbeke and Molenberghs 2000, chapter 9) has been followed: first we fit a standard OLS regression to our data using all available covariates; then we analyse the resulting residuals as a guide to identifying the random effects. This preliminary analysis has been conducted separately on several random samples of size 1,000, drawn up according to the replication design described above.

The adjusted R^2 of the OLS regression is close to 0.35 in every observed sample. This rather low figure is the result of the nature of the phenomenon under study (household income is not easy to predict), the information contained in the survey and the constraint represented by the need to include only those covariates for which the population total can be obtained from the Census.

Figure 1 contains “box and whiskers” plots of the residuals by area and wave constructed for one of the Monte Carlo sample (very similar findings may be observed in every sample). Analysis of the plots suggests that there is within-area and within-wave correlation, and thus the need to specify models including area and wave effects. From an analysis of residuals, it is less clear whether the inclusion of interaction effects (that is time varying area effects) would be beneficial or not.

Moreover the residuals show a degree of autocorrelation, the average of the autocorrelation coefficient calculated over all individual residual histories being 0.27. Even though this autocorrelation level is not very high, for the sake of completeness we decided to also take into consideration models with autocorrelated errors or random effects. After having tested various different autocorrelation structures (ARMA(p, q), General Linear, *etc.*), we found that the autoregressive process of order 1 provides the best fit to our data.

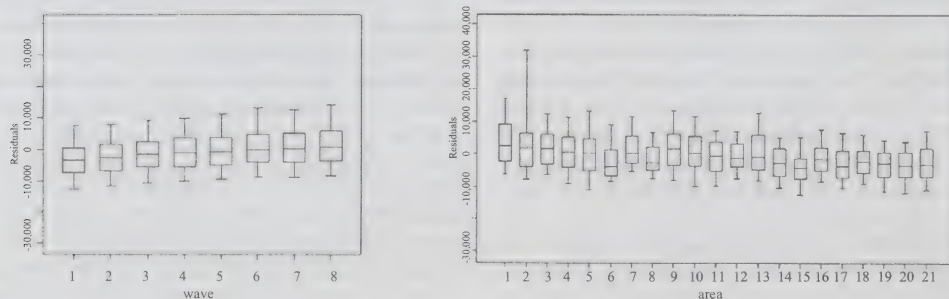


Figure 1 Box and whiskers plot of residuals by wave (left) and area (right)

The apparent skewness of residuals also suggests that the normality assumption for errors does not hold exactly. We maintain this assumption for all the models we specify, and we use REML estimators for variance components. In fact, we may expect departures from normality to have a slight impact on point values of predictors. BLUP formulas can be derived without normality; moreover, there are sound reasons for us to expect REML (and ML) estimators of ψ to perform well even if normality does not hold (see Jiang, 1996, for details). Departures from normality may have more a serious impact on MSE estimation, and this is a problem we are going to be looking at in section 4.2 below.

4. Results

4.1 Point estimators

All computations involved in the simulation exercise described in section 3 were carried out using SAS version 9.1 for Windows. EBLUP estimators are obtained using Proc MIXED, and the generation of samples is based on Proc SURVEYSELECT.

Given that the primary goal of Small Area Estimation is the precise estimation of area-specific parameters, we first evaluated how well the described estimators perform when predicting individual area values. Moreover, we also evaluated the amount of over-shrinkage connected with each estimator. In fact, small area estimates should reflect (at least approximately) the variability in the underlying area parameters taken as a whole.

We note that our simulation experiment is aimed at evaluating design-based properties of the estimators, that is, the population from which the random samples are generated is held fixed.

For the evaluation of the estimators' performance, we adopted an approach that is commonly found in the literature (see Rao 2003; section 7.2.6), using two indicators, the Average Absolute Relative Bias (AARB) and the Average Relative Mean Square Error (ARMSE):

$$\begin{aligned} \text{AARB} &= m^{-1} \sum_{d=1}^m \left| R^{-1} \sum_{r=1}^R \left(\frac{\tilde{y}_{dT(r)}}{\bar{Y}_{dT}} - 1 \right) \right| \\ \text{ARMSE} &= m^{-1} \sum_{d=1}^m \left\{ R^{-1} \sum_{r=1}^R \left(\frac{\tilde{y}_{dT(r)}}{\bar{Y}_{dT}} - 1 \right)^2 \right\} \end{aligned} \quad (18)$$

where $\tilde{y}_{dT(r)}$ is the estimate for area d , time T and replicated sample r , while \bar{Y}_{dT} is the population mean being estimated. Note that AARB measures the bias of an estimator, whereas ARMSE measures its accuracy. The number of replications R is set at 500, a figure large enough to obtain stable Monte

Carlo estimates of expected values and variances, frequently used in simulation studies on small area estimation (Heady, Higgins and Ralphs 2004; EURAREA Consortium 2004).

The gain in efficiency connected to each small area estimator is evaluated using the ratio of its ARMSE to the ARMSE of certain estimators we use as benchmarks. In particular, all estimators are compared with the weighted mean $\bar{y}_{dT, \text{DIR}}$ and we denote this ratio as AEFF_{Dir} . Moreover, EBLUP estimators associated with models (5) - (9), which use data from previous waves, are compared with the EBLUP estimator associated with the cross-sectional model (10), in order to assess the gain in efficiency deriving from the use of past waves. In this case the ratio is denoted as $\text{AEFF}_{\text{Sect}}$.

As far as the evaluation of the degree of shrinkage is concerned, we have compared the empirical standard deviation of population area values:

$$\text{ESD} = \sqrt{m^{-1} \sum_{d=1}^m (\bar{Y}_{dT} - \bar{Y}_T)^2},$$

where \bar{Y}_T is the mean of the population values of the m areas at time T , with the empirical standard deviation of the estimated area values, which in the case of a simulation study is given by:

$$\text{esd} = R^{-1} \sum_{r=1}^R \left[\sqrt{m^{-1} \sum_{d=1}^m (\bar{y}_{dT(r)} - \bar{\bar{y}}_{T(r)})^2} \right],$$

where $\bar{\bar{y}}_{T(r)}$ is the mean of the estimated values for the m areas at time T in the simulation run r . The comparison is carried out using the indicator

$$\text{RESD} = \frac{\text{esd}}{\text{ESD}} - 1 \quad (19)$$

which tells us how the empirical standard deviation associated with one estimator differs from that of the population.

Table 1 contains the percentage values of AARB, ARMSE, AEFF and RESD obtained for the direct estimator, the design-based estimators given in (2) and (3) and the EBLUP estimators derived from models (5) - (10).

All estimators perform significantly better than $\bar{y}_{dT, \text{DIR}}$ in terms of ARMSE, leading to less than 100% AEFF_{Dir} values. We can also see that design-based estimators are worse than EBLUP estimators in terms of ARMSE, and that the gain in efficiency demonstrated by AEFF_{Dir} is particularly high in some cases (in excess of 50%). This result highlights the superior accuracy of the model-based estimators in question.

Table 1 Performance indicators - auxiliary information is available

Model	AARB%	ARMSE%	AEFF _{Dir} %	AEFF _{Sect} %	RESD%
DIR	0.0	0.787	100.0	-	15.6
COMP	2.7	0.552	70.1	-	-9.8
GREG	0.2	0.543	68.2	-	10.0
SMM	2.3	0.377	47.7	100.0	-8.7
MM1	3.1	0.358	45.3	95.0	2.4
MM2	2.4	0.427	54.1	113.4	-4.4
MM3	2.6	0.380	48.3	101.2	4.7
MM4	2.6	0.429	54.2	113.6	-8.0
MM5	2.9	0.318	40.4	84.7	-7.7

The most reliable EBLUP estimator is the one associated with the *MM5* model, with independent area and time effects and residuals autocorrelated according to an AR(1) process, leading to a gain in efficiency of about 60% compared with the direct estimator. This is followed by the EBLUP estimator associated with model *MM1*, which differs from the previous one only because of the absence of autocorrelated residuals.

In terms of bias, the GREG estimator gives the smallest value of AARB, as would be expected (Särndal, Swensson and Wretman 1992, chapter 7; Veijanen, Lehtonen and Särndal 2005). This is followed by the remaining estimators, all of which reveal a similar value for AARB. Of the EBLUP estimators, those associated with the *MM1* and *MM5* models are more efficient in terms of ARMSE, but they are slightly more biased than the one associated with the *SMM*. This is probably due to the fact that we limit our evaluation of performance to the last wave; for this data subset we would expect the fit of the regression underlying *SMM*, based on the last wave only, to be better than the one based on the whole data set. As far as EBLUP estimators are concerned, the $AEFF_{Sect}$ column shows how the gain in efficiency of the predictors, based on borrowing strength over time, is positive in some cases and negative in others. Models *MM2* and *MM4* (see formulas (6) and (8)), where effects of interaction between area and time are present, are apparently inadequate because the predictors associated with both models perform rather poorly. The performance of the predictor associated with *MM3* (see (7)) is also slightly worse than that of the predictor associated with the cross-sectional model: this rather surprising result is probably due to the low number of waves, which does not allow for an effective estimation of the correlation coefficient between consecutive time effects.

As we have already said, the estimator associated with model *MM5* is the one that performs the best: it is considerably more efficient than the one associated with *SMM*, with an $AEFF_{Sect}$ of roughly 85% representing a gain in efficiency of about 15% due to consideration of more than one wave. The EBLUP estimator associated with *MM1* also turns out to be more efficient than the one associated with *SMM*, but in this case the gain is one of only 5%.

These results confirm the fact that household level data at several consecutive points in time may be employed, via certain kinds of longitudinal model, to produce more efficient estimates.

Moving on to the indicator for shrinkage reported in the last column of the table, we can see that the direct estimator overestimate the standard deviation of the population of area means, by 15%. The same effect, albeit somewhat attenuated, is observed for the GREG estimator, whose standard deviation is over-inflated by 10%. On the contrary, the COMP estimator tends to “shrink” the estimates towards the centre of the distribution, leading to a reduction in the standard deviation of area means of about 10% with respect to the population. These results are in line with those obtained by other authors comparing the same kinds of estimator (Heady *et al.* 2004; Spjøtvoll and Thomsen 1987). The results obtained for EBLUP estimators are more encouraging, as the calculated percentage difference is always less than 10% in absolute terms. Hence, in this respect all EBLUP estimators seem to be acceptable. Moreover, we may expect that the BLUP estimators are under-dispersed compared to the corresponding population parameters. In this case, the indicator RESD assumes positive values for some longitudinal EBLUP estimators because it is calculated only on the last wave, while longitudinal models are aimed to predict $m \times T$ parameters.

Table 2 summarizes the results regarding those EBLUP estimators associated with random error variance models, as described in the last paragraph of section 2. When no auxiliary variables are included in the models, the advantage of “borrowing strength” over time and area is singled out independently of the advantage associated with covariates.

As expected, the improvements in efficiency measured by $AEFF_{Dir}$ are smaller than those shown in Table 1, although the reductions in ARMSE remain significant. The ranking of those predictors associated with the various random effects specification remains the same as the one presented in Table 1, the predictor associated with the *MM5** model resulting the most efficient, as shown by ARMSE%. The gain in efficiency associated with this latter estimator compared with the direct estimator is about 43%.

Table 2 Performance indicators - auxiliary information is unavailable

Model	AARB%	ARMSE%	AEFF _{Dir} %	AEFF _{Sect} %	RESD%
SMM*	2.7	0.575	72.8	100.0	-7.6
MM1*	2.9	0.556	70.3	96.6	7.5
MM2*	2.8	0.639	80.8	111.0	-3.0
MM3*	3.7	0.574	72.6	99.7	8.6
MM4*	3.5	0.691	87.2	119.8	-6.7
MM5*	3.0	0.445	56.2	77.2	-6.3

With regard to bias, the EBLUP estimators obtained from those models with no covariates tend to be more biased than the corresponding ones with covariates.

The analysis of the AEFF_{Sect} column shows that the reduction in ARMSE allowed for by some of those models borrowing strength over time, is larger than in the case where covariates are included, as it reaches 22% in the best example of the MM5* model.

This last result is really encouraging. In fact, within the context of Small-Area Estimation, the absence of any known totals of covariates in the population can be very limiting when trying to obtain reliable estimates. The observed ARMSE reduction connected to the consideration of more waves in a panel survey show that estimates may be improved “borrowing strength” over time, when it is not possible to exploit auxiliary information.

With regard to the results of shrinkage, they may be considered acceptable also in this case, and one can see a relationship between the results obtained for EBLUP estimators derived from analogous models with or without covariates.

4.2 Comparing different estimators of the MSE of EBLUP estimators

In section 2.3 we reviewed three different estimators of the MSE associated with EBLUP estimators. In this section we are going to compare the performances of these three estimators using our simulation exercise. Given that we are focusing on MSE estimation rather than a comparison of EBLUP estimators derived from different models, we only consider the predictor associated with model MM5, which emerged as the best performer in the previous section.

Let us denote the predictor of \bar{Y}_{dt} with $\hat{\eta}_{dt}^{EBLUP}$ and its mean square error as $MSE(\hat{\eta}_{dt}^{EBLUP})$. The following estimator:

$$mse_{ACT}(\hat{\eta}_{dt}^{EBLUP}) = \frac{1}{R} \sum_{r=1}^R (\hat{\eta}_{dt(r)}^{EBLUP} - \hat{\eta}_{dt}^{EBLUP})^2 + (\hat{\eta}_{dt}^{EBLUP} - \bar{Y}_{dt})^2,$$

where $\hat{\eta}_{dt(r)}^{EBLUP}$ is $\hat{\eta}_{dt}^{EBLUP}$ calculated on the r^{th} replicated sample and $\hat{\eta}_{dt}^{EBLUP} = R^{-1} \sum_{r=1}^R \hat{\eta}_{dt(r)}^{EBLUP}$, will be used as benchmark for the comparison of the performance of the mean square error estimators described in section 2.3, because the true mean squared error is not known.

As in the case of point estimators, all computations are done using SAS. To determine the Prasad-Rao estimator (14), the output of Proc MIXED's ESTIMATE statement is used with the option KENWARDROGER activated. The sum $g_1(\hat{\psi}) + g_2(\hat{\psi})$ is obtained from the output of Proc MIXED. The KENWARDROGER option allows for the calculation of an MSE inflation factor, described in Kenward and Rogers (1986), which is equivalent to $2g_3(\hat{\psi})$ (see also Rao 2003, section 6.2.7).

The estimator $mse_{BL}(\hat{\eta}_{dt}^{EBLUP})$ is re-sampling based. Hence the evaluation of its performance with respect to a Monte Carlo exercise requires the implementation of two nested simulations: for each $r (r=1, \dots, R)$, we run the R_{BOOT} replications needed to approximate expectations with respect to the bootstrap model. To limit the computational burden, we set $R_{BOOT}=150$. Butar and Lahiri (2003) propose an analytical approximation of mse_{BL} , but only for models that are not as complex as the one in question.

For both $mse_{BL}(\hat{\eta}_{dt}^{EBLUP})$ and $mse_{JLW}(\hat{\eta}_{dt}^{EBLUP})$, we have prepared *ad-hoc* SAS codes using the output of Proc MIXED as inputs.

In order to compare the three MSE estimators, we employ the same measures used to evaluate the performance of point estimators, AARB and ARMSE. As there is usually some concern about the under-estimation of MSE estimators, we are also interested in the sign of any bias associated with the estimators in question. Therefore, in the case of MSE estimators we do not only calculate the average of the absolute values of the estimates obtained for the bias in each region (AARB), but also the average of these estimates without the absolute value (AARB*), so as to better understand whether the given estimators indeed tend to under-evaluate the MSE or not. Hence the calculated measures are:

$$\begin{aligned} \text{AARB} &= m^{-1} \sum_{d=1}^m \left| R^{-1} \sum_{r=1}^R \left(\frac{\text{mse}_*(\tilde{\eta}_{dt}^{\text{EBLUP}})}{\text{mse}_{\text{ACT}}(\tilde{\eta}_{dt}^{\text{EBLUP}})} - 1 \right) \right| \\ \text{AARB}' &= m^{-1} \sum_{d=1}^m \left\{ R^{-1} \sum_{r=1}^R \left(\frac{\text{mse}_*(\tilde{\eta}_{dt}^{\text{EBLUP}})}{\text{mse}_{\text{ACT}}(\tilde{\eta}_{dt}^{\text{EBLUP}})} - 1 \right) \right\}, \\ \text{ARMSE} &= m^{-1} \sum_{d=1}^m \left\{ R^{-1} \sum_{r=1}^R \left(\frac{\text{mse}_*(\tilde{\eta}_{dt}^{\text{EBLUP}})}{\text{mse}_{\text{ACT}}(\tilde{\eta}_{dt}^{\text{EBLUP}})} - 1 \right)^2 \right\} \end{aligned}$$

where the symbol * refers to the considered estimation procedures, that are PR, BL, JLW. Results of the comparisons based on $R = 500$ MC iterations are reported in Table 3.

Table 3 Performance of MSE estimators of $\tilde{\eta}_{dt}^{\text{EBLUP}}$ under model MM5

Estimator	AARB	AARB'	ARMSE
mse _{PR}	0.378	-0.383	0.238
mse _{BL}	0.377	-0.318	0.228
mse _{JLW}	0.337	0.036	0.261

In terms of ARMSE and AARB, the three estimators behave similarly, with no particular one emerging as clearly better than the other two. Nonetheless, the AARB' column clearly shows that mse_{PR} and mse_{BL} systematically underestimate MSE_{ACT}, whereas mse_{JLW} does not. This is probably due to the failure of the normality assumption for error terms. In fact, as we foresaw in section 3, equalized income is a positively skewed variable, and the regression residuals \mathbf{e} also appear to be so. Normality is a crucial assumption in the derivation of mse_{PR} and mse_{BL}, while mse_{JLW} could be expected to be more robust in this respect. Our findings are consistent with the theory predictions and simulation results described in Jiang *et al.* (2002). Although Bell (2001) noted that mse_{JLW} may be negative for some data set because of the bias correction, this never happens in our simulations. For all replicated data set we have that the second term in (17) gives a positive, and in most cases substantial contribution to the estimate of the MSE. A discussion of modifications of (17) when it returns negative values can be found in Jiang and Lahiri (2006b).

To conclude then, in the case of the present problem, mse_{JLW} emerges as the most appropriate of the three measures for estimating MSE($\tilde{\eta}_{dt}^{\text{EBLUP}}$). This finding could be of importance for any application of normality-based linear mixed models theory to data set in which normality assumptions for error terms do not hold exactly.

We replicated the simulation exercise also for the cross-section model without covariates SMM^* , that is often considered in simulations aimed at the comparison of different estimation methods. To this end we note that for this model the ratio $\hat{\sigma}_e^2/\hat{\sigma}_v^2$ is around 12, leading to a

EBLUP predictor characterized by $\gamma_i = \hat{\sigma}_v^2 n_i / (\hat{\sigma}_v^2 n_i + \hat{\sigma}_e^2)^{-1}$ ranging from 0.54 to 0.9. We note also that some, but not all, areas are characterized by the presence of outliers (skewness coefficient γ_1 ranges from 0.1 to 4.6).

In this setting MSE estimators show a behavior quite different form that illustrated in the case of model MM5. Results are shown in Table 4.

Table 4 Performance of MSE estimators of $\tilde{\eta}_{dt}^{\text{EBLUP}}$ under model SSM*

Estimator	AARB	AARB'	ARMSE
mse _{PR}	0.449	0.262	0.503
mse _{BL}	0.376	0.213	0.376
mse _{JLW}	0.354	0.149	0.335

All estimators overestimate the actual MSE, although mse_{JLW} overestimates less than the other two. From a detailed analysis of results related to individual areas, we have the values of AARB' (that represents the most apparent difference with the results of Table 3) is driven by severe overestimation of actual MSE in areas characterized by the lowest levels of skewness and kurtosis. For these areas $\hat{\sigma}_e^2$ largely overstates actual variation in the data, thus leading to overestimation of $g_1(\sigma_v^2, \sigma_e^2)$. This is likely to be due to the fact that the failure of normality (the excess of kurtosis) causes the overestimation of σ_e^2 . This problem did not appear in the case of model MM5 because of the presence of covariates and the AR(1) modeling of individual residuals.

5. Concluding remarks and further developments

The results obtained show that, in general, EBLUP estimators derived from unit level linear mixed model specifications that “borrow strength over time”, as well as over areas, provide a significant gain in efficiency compared with both the direct estimator and with other commonly-used design based estimators such as the optimal composite estimator and the GREG estimator. Moreover, the mean squared error of some of the longitudinal EBLUP estimators in question is considerably lower, on average over the areas, than that of the analogous cross-sectional EBLUP estimators. Among the model specifications used to derive EBLUP estimators, those with independent time and area effects, whether inclusive of the autocorrelation of residuals or not, appear the most efficient, offering a gain in efficiency of about 55-60% compared with the direct estimator. These results also hold when covariates are removed; in fact, they offer the chance to obtain reliable small area estimates even in the absence of covariates, provided that repeated observations of the same unit at several points in time are available. Besides the shrinkage

effect connected to EBLUP estimators appears moderate, reducing the need for ensemble or multiple estimation (Rao 2003, Chapter 9). With regard to estimation of the MSE of the small area estimators in question, we noted that the jackknife estimator provides the best results being correct, on average, over the areas and thus more robust to any departure from the standard assumptions of linear mixed models. This finding may be of importance to all applications of normality-based linear mixed models theory to data set in which normality assumptions do not hold exactly, as in the case of income data.

Acknowledgements

Research was partially funded by Miur-PRIN 2003 "Statistical analysis of changes of the Italian productive sectors and their territorial structure", coordinator Prof. C. Filippucci. The work of Enrico Fabrizi was partially supported by the grants 60FABR06 and 60BIFF04, University of Bergamo.

We thank ISTAT for kindly providing the data used in this work.

References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Bell, W. (2001). Discussion with "Jackknife in the Fay-Herriott model with an example", *Proceeding of the Seminar on Funding Opportunity in Survey Research*, 98-104.
- Betti, G., and Verma, V. (2002). Non-monetary or lifestyle deprivation, in EUROSTAT (2002). *Income, Poverty Risk and Social Exclusion in the European Union*, Second European Social Report, 87-106.
- Butar, F., and Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76.
- Datta, G.S., and Lahiri, P. (1999). A unified measure of uncertainty of estimated best linear predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical bayes estimation of unemployment rates for the States of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- EURAREA CONSORTIUM (2004). EURAREA. Enhancing Small Area Estimation Techniques to meet European Needs, Project Reference Volume, downloadable at <http://www.statistics.gov.uk/eurarea/download.asp>.
- EUROSTAT (2002). European social statistics - Income, poverty and social exclusion. 2nd report.
- Falorsi, P.D., Falorsi, S. and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology*, 20, 171-176.
- Falorsi, P.D., Falorsi, S. and Russo, A. (1999). Small area estimation at provincial level in the Italian Labour Force Survey. *Journal of the Italian Statistical Society*, 1, 93-109.
- Ghosh, M., Nangia, N. and Kim, D. (1996). Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Heady, P., Higgins, N. and Ralphs, M. (2004). Evidence-Based Guidance on the Applicability of Small Area Estimation Techniques. Paper presented at the European Conference on Quality and Methodology in Official Statistics, Mainz, Germany, May, 24-26.
- Jiang, J. (1996). REML estimation: Asymptotic Behavior and Related Topics. *The Annals of Statistics*, 24, 255-286.
- Jiang, J., Lahiri, P. and Wan, S.M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.
- Jiang, J., and Lahiri, P. (2006a). Estimation of finite population domain means - A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.
- Jiang, J., and Lahiri, P. (2006b). Mixed model prediction and small area estimation. Editor's invited discussion paper, *Test*, 15, 1, 1-96.
- Kenward, M.G., and Roger, J.H. (1986). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.
- Lehtonen R., Särndal C.-E. and Veijanen A. (2003) The effect of model choice in estimation for domains, including small domains, *Survey Methodology*, 29, 1, 33-44.
- Pfaffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review*, 70, 125-143.
- Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.
- Rao, J.N.K. (2003). *Small Area Estimation*, New York: John Wiley & Sons, Inc.
- Rao, J.N.K., and You, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.
- Rendtel, U., Behr, A. and Sisto, J. (2003). Attrition effect in the European Community household panel, CHINTX PROJECT, European Commission.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Singh, A.C., Mantel, H.J. and Thomas, B.W. (1994). Time series EBLUPS for small areas using survey data. *Survey Methodology*, 20, 1, 33-43.
- Spjøtvoll, E., and Thomsen, I. (1987). Application of some empirical bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, 4, 435-450.

- Vandecasteele, L., and Debels, A. (2004). Modelling attrition in the European Community Household Panel: The effectiveness of weighting, 2nd International Conference of ECHP Users, EPUNet 2004, Berlin, June 24-26.
- Verbeke, G., and Molenberg, H. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.
- Vejjanen, A., Lehtonen, R. and Särndal, C.-E. (2005). The Effect of Model Quality on Model-Assisted and Model-Dependent Estimators of Totals and Class Frequency for Domains, paper presented at SAE2005 Conference, Challenges in Statistics Production for Domains and Small Areas, August, 28-31 2005, Jyväskylä, Finland.

Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results

Anne Renaud¹

Abstract

Coverage deficiencies are estimated and analysed for the 2000 population census in Switzerland. For the undercoverage component, the estimation is based on a sample independent of the census and a match with the census. For the overcoverage component, the estimation is based on a sample drawn from the census list and a match with the rest of the census. The over- and undercoverage components are then combined to obtain an estimate of the resulting net coverage. This estimate is based on a capture-recapture model, named the dual system, combined with a synthetic model. The estimators are calculated for the full population and different subgroups, with a variance estimated by a stratified jackknife. The coverage analyses are supplemented by a study of matches between the independent sample and the census in order to determine potential errors of measurement and location in the census data.

Key Words: Census; Coverage errors; Dual system; Multi-stage sampling plan; Measurement errors.

1. Introduction

In any census, some people are not enumerated and should be, while others are counted twice or should not have been enumerated. There is both undercoverage and overcoverage, and quite often, the combined result is net undercoverage. For example, net undercoverage is estimated at 1.6% in the United States in 1990 (Hogan 1993), 2.2% in the United Kingdom in 1991 (Brown, Diamond, Chambers, Buckner and Teague 1999) and 3% in Canada in 2001 (Statistics Canada 2004). By contrast, in the United States in 2000, there is estimated to be net overcoverage of 0.5% (Hogan 2003). Coverage deficiencies may vary greatly between subgroups of the population. In the United States in 2000, blacks were found to have a net undercoverage of 1.8%, while whites had an overcoverage of 1.1%. Also, values often vary between age classes and regions, for example. These coverage deficiencies, and other errors such as measurement errors, result in a biased picture of the population. They are therefore studied in order to obtain information on the quality of the available data and to find ways to improve censuses of the population.

The 2000 population census in Switzerland gives a picture of the population on December 5, 2000. In this article, coverage deficiencies in a Swiss census are estimated for the first time. Undercoverage, overcoverage and net coverage resulting from the 2000 census are all analysed. Undercoverage is estimated from a sample of individuals S_p , independent of the census, on which a coverage survey was organized a few months after the census (collection took place in April and May 2001). The data from the survey are matched with data from the census to determine whether persons in S_p were enumerated.

Overcoverage is estimated from a sample of individuals S_E drawn from census records. A search for duplicates and other erroneous records then serves to determine whether a given record corresponds to a real person to be enumerated. Net coverage is estimated on the basis of a capture-recapture model known as the dual system (Wolter 1986, Fienberg 1992). The dual estimator is applied in homogeneous cells, and the results are recombined using a synthetic model to obtain results for different domains of the population (Hogan 2003). The purpose of the project is not to adjust the census figures but rather to obtain information on the quality of the 2000 census and potential improvements for future censuses.

This article describes the different steps followed in obtaining estimates, then presents the results. Sections 2 and 3 describe the data sets and the coverage estimators. Section 4 provides the details on constructing the different statuses used in the estimators. Section 5 describes the approach used to compare the values collected in the census and in the survey for the matched persons from S_p . Sections 6 and 7 present the numerical results and the conclusion.

2. The three data sets

2.1 Census

The 2000 census was conducted under the auspices of the Federal Statistical Office, with the reference date of December 5, 2000. Information was collected for 7.3 million inhabitants, 3.1 million households, 3.8 million dwellings and 1.5 million buildings. The different levels were then linked by common identifiers when the data were processed.

1. Anne Renaud, Service de méthodes statistiques, Office fédéral de la statistique, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland. E-mail: Anne.Renaud@bfs.admin.ch.

The collection of information on persons and households was the responsibility of Switzerland's 2,896 political communes. The latter had a choice between different methods of collection:

- TRADITIONAL: use of census agents;
- SEMI-TRADITIONAL: pre-printed questionnaires based on the communal register of inhabitants are mailed out and then collected by census agents;
- TRANSIT: pre-printed questionnaires are mailed out and mailed back;
- FUTURE: identical to TRANSIT except that links between households and dwellings are supplied by the commune;
- TICINO: similar to TRANSIT but limited to the canton of Tessin.

Most of the SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes also offered the option of completing questionnaires online. The 2,208 SEMI-TRADITIONAL, TRANSIT, FUTURE and TICINO communes that used the pre-printing of questionnaires based on communal registers of inhabitants account for nearly 96% of the population. For most of these communes, the tasks of mailing out questionnaires and controlling their return were organized at a national centre.

The data set for individuals contains 7,452,075 entries. One feature of this data set is that it contains two records for the same person if that person has two residences (2.3% of the population; for example, a student who both resides with his parents and has a residence close to his school). In the case of two residences, one is coded as the *economic residence* and the other as the *civil residence*. The economic residence is the place where the person spends the most time per week and the civil residence is where the person's official papers are kept (birth certificate for Swiss citizens, residence permit for foreigners). Where there is just one residence, it is both the economic and the civil residence. Switzerland is considered to have a *resident population* of 7,280,010 based, on the set of records showing the economic residence.

Households are classified as *private*, *collective* or *administrative*. Examples of private households are families, couples and persons living alone. Examples of collective households are groups of occupants of a home for the aged or a boarding school or the inmates of a prison. Administrative households group together people with no fixed residence, travellers and persons - by building or commune - who could not be assigned to private or collective households (2.4% of the resident population).

Census data contain no imputation at the record level, since communes sent basic information for non-respondents (unit non-response). However, values are imputed in the

case of missing data or inconsistency in questionnaires (item non-response).

The *population of interest* for coverage estimates is the resident population (based on economic residence) in private and administrative households. Collective households, which account for 2.3% of the enumerated resident population, are excluded from the estimates.

2.2 S_p sample, coverage survey and matching (undercoverage)

The size objective for the S_p sample is set at approximately 50,000 people. In the absence of existing frames in Switzerland, this value was determined approximately, based on experiences in other countries. In particular, the Australian results for 1996 were used, since the sampling plan for Australia's coverage survey was similar to the one for Switzerland in 2000 (ABS 1997).

The S_p sample, which is independent of the census, is constructed in two parts: the canton of Tessin (TICINO) and the rest of Switzerland (NORD). Both parts use a multi-stage draw. The first stage consists in selecting 303 primary units - these are political communes for TICINO and postal codes for NORD - according to a stratified plan with a draw proportional to the number of buildings. The second stage consists in a simple random draw of a fixed number of 60 buildings per primary unit. In the NORD plan, these buildings are allocated to a maximum of three mail delivery routes, based on an intermediate sampling stage. The sampling is thus constructed so as to consolidate the field work while limiting the variability of the weights. For practical reasons and in light of available resources, postal codes that include a large proportion of buildings lacking complete postal addresses or coded as unoccupied are selected with a lower probability than other postal codes. These tend mainly to be postal codes in rural areas or industrial zones, which are unlikely to exhibit major coverage deficiencies. With the assistance of postal employees, complete lists of households are drawn up in the field within the sample of approximately 16,000 buildings. A sub-sample of buildings is then drawn so as to obtain a total of approximately 27,000 households. For more information on the sampling and survey procedure, see Renaud (2001) and, in greater detail, Renaud and Eichenberger (2002).

The coverage survey consists in contacting the 27,000 households - by telephone if a telephone number is found and in person if not. The variables collected are those that lend themselves to matching with the census and defining subgroups of interest for the coverage study (socio-demographic variables, addresses). The collection operation covers all members of all households in the selected buildings.

The final sample S_p contains $n_p = 49,883$ people in the population of interest (persons listed at their economic residence and residing in a private household). Of the households contacted, 88% were reached by telephone and 12% in person. The weighting depended on the sampling and an adjustment for non-response. The adjustment for non-response was based on a homogeneity model in cells constructed on the basis of the sampling strata and whether or not a telephone number was known to exist (interviews conducted by telephone or in person). It also incorporated an estimate of the proportion of true households among the households to be contacted, since a sizable portion of the households to be contacted actually consisted of vacant dwellings, stores or businesses. No calibration was applied, since the auxiliary data available were not independent of the census. There was no partial non-response. The weighting details are documented in Renaud and Poterat (2004).

Based on the questions asked in the survey and various plausibility controls, we hypothesize that the S_p data are correct and usable for matching with the census. The quality criteria used are as follows:

- *completeness*: the record is sufficient to identify the person;
- *appropriateness*: the person should have been enumerated;
- *uniqueness*: the person is listed only once;
- *belonging to population of interest*: the person is listed at his/her economic residence and in a private household;
- *correctness of location*: the person is listed at the correct address on Census Day.

The matching between the S_p sample and the census serves to determine the *matching status* P_j of each element j of S_p . Status P_j is equal to 1 if the element is matched in the census (enumerated person) and 0 if this is not the case (person not enumerated). In our case, the data collected in the coverage survey, the final census data and images of the census questionnaires are used for automatic matching, manual matching and controls. No supplementary interview took place in addition to the coverage survey. Persons who moved between Census Day and the day of the survey were sampled at their address on the day of the survey and then searched for on a priority basis at the address they had indicated for the day of the census. No case was unresolved by the end of the process.

2.3 S_E sample and search for erroneous records (overcoverage)

The size objective for the S_E sample was set at approximately 55,000 persons. This value, somewhat greater than

that of S_p , had little influence on the processing of the data, since there was no field work or interview supplementary to the census.

The S_E sample was selected from the census data using a two-stage draw. Only elements included in the population of interest were eligible (records at the economic residence, without members of collective households). The primary units of S_E were identical to the primary units of S_p (postal codes and communes). However, the list of postal codes in the NORD plan that was used for S_p did not correspond exactly to the list of postal codes that were present in the census data. Census records found in postal codes that did not exist in the list used for S_p were therefore reallocated to existing codes, taking geographic location into account (this involved assigning fictitious postal codes for the sampling). In the second stage, records were drawn from the population of interest using a simple random plan, without intermediate stages. The allocation was done in such a way as to obtain constant weights in the sampling strata of the primary units. In the end, the sample contained $n_E = 55,375$ records (Renaud 2003).

We hypothesize that S_E records are sufficient to identify persons (completeness), since there is little imputation in the census data and most questionnaires were pre-printed based on registers of inhabitants. Appropriateness and uniqueness were determined in a matching between S_E and the rest of the census using a procedure similar to the matching between S_p and the census. In our case, this involves a search for duplicates or triplicates of elements of S_E , supplemented by an analysis of suspect cases in S_E . An element j is considered appropriate if it is not considered erroneous in the analysis of suspect cases (e.g., a note on the questionnaire indicating that the person has gone abroad). An element j is considered unique if no duplicate or triplicate is detected in the census. There is no supplementary interview for S_E . There is therefore no information supplementary to the census for S_E persons (actual location? actual type of residence or household?). The search for duplicates/triplicates and suspect cases results in an *enumeration status* E_i for each element j of S_E . Status E_i is equal to 1 if the element should indeed have been enumerated in the census (default value) and 0 if it should not have been enumerated. In practice, it can take on values between 0 and 1 if the case is not determined precisely. Thus, duplicates and triplicates receive respectively the values 1/2 and 1/3 if there is no information allowing the correct record to be determined from among the records detected. These cases, which are rare, consist of persons who completed more than one questionnaire in the census without any link having been made between those questionnaires during the processing of the data.

3. Coverage estimators

3.1 Undercoverage and overcoverage

The *undercoverage rate* is estimated by $\hat{R}_{\text{under}} = 1 - \hat{R}_m$, where \hat{R}_m is the estimate of the *correct matches rate* based on the S_p sample. Similarly, the *overcoverage rate* is defined as $\hat{R}_{\text{over}} = 1 - \hat{R}_c$, where \hat{R}_c is the estimate of the *correct records rate* based on the S_E sample. The correct matches rate and the correct records rate are estimated by the weighted means of matching status P_j and enumeration status E_j , as follows:

$$\hat{R}_m = \frac{\sum_{j \in S_p} w_{p,j} P_j}{\sum_{j \in S_p} w_{p,j}} \quad \text{and} \quad \hat{R}_c = \frac{\sum_{j \in S_E} w_{E,j} E_j}{\sum_{j \in S_E} w_{E,j}}, \quad (1)$$

where $w_{p,j}$ is the weight of element j of sample S_p and $w_{E,j}$ is the weight of element j of sample S_E . We note that the denominator of \hat{R}_c is the sum of the weights $w_{E,j}$ of S_E and not the number C of known records in the census, so as to have a potentially less biased estimator.

The estimate of the undercoverage and overcoverage rates in a domain d is given by $\hat{R}_{\text{under},d} = 1 - \hat{R}_{m,d}$ and $\hat{R}_{\text{over},d} = 1 - \hat{R}_{c,d}$, with

$$\hat{R}_{m,d} = \frac{\sum_{j \in S_p} w_{p,j} P_j I_{jd}}{\sum_{j \in S_p} w_{p,j} I_{jd}} \quad \text{and} \quad \hat{R}_{c,d} = \frac{\sum_{j \in S_E} w_{E,j} E_j J_{jd}}{\sum_{j \in S_E} w_{E,j} J_{jd}}. \quad (2)$$

Identifiers I_{jd} and J_{jd} take on the value 1 if element j , respectively of S_p and S_E , is found in domain d ; otherwise their value is 0.

3.2 Net coverage

The *net undercoverage rate* is estimated by $\hat{R}_{\text{netunder}} = 1 - \hat{R}_{\text{net}}$ where $\hat{R}_{\text{net}} = C / \hat{N}$ is the estimate of the *net coverage rate*, C is the number enumerated in the population of interest and \hat{N} is the estimate of the true total in the population of interest. If $\hat{R}_{\text{netunder}}$ is negative, there is net overcoverage.

The estimate of the true total \hat{N} is based on the dual model (Wolter 1986). This model is built on the principle of capture (census) and recapture (coverage survey). It is applied in estimation cells $k = 1, \dots, K$ in order to best satisfy the assumptions of the model; see discussion below. Thus, the estimate of the true total \hat{N} is composed of the sum of the estimated true totals \hat{N}_k in disjoint estimation cells covering the population of interest $k = 1, \dots, K$:

$$\hat{N} = \sum_{k=1}^K \hat{N}_k. \quad (3)$$

The estimated totals \hat{N}_k have the form given by the dual model:

$$\hat{N}_k = [N_{1+,k}] \begin{bmatrix} N_{+,k} \\ N_{11,k} \end{bmatrix}, \quad (4)$$

where $N_{1+,k}$ is the total of records correctly counted in cell k during capture (census), $N_{+,k}$ is the total in k during recapture (estimated from sample S_p) and $N_{11,k}$ is the number of records common to the two lists (estimated from matches between S_p and the census).

The different terms of equation (4) are estimated using undercoverage and overcoverage estimates. This is an extension of the model in Wolter (1986), similar to the one used by Hogan (2003). Thus, the total of the records correctly counted in the census $N_{1+,k}$ is estimated by the enumerated total C_k multiplied by the correct records rate $\hat{R}_{c,k}$ to take account of overcoverage. Also, the ratio between the total in the recapture $N_{+,k}$ and the number of records common to the two lists $N_{11,k}$ is estimated by the inverse of the rate of matching $\hat{R}_{m,k}$ between the coverage survey and the census in order to take account of undercoverage. We obtain

$$\hat{N}_k = [C_k \hat{R}_{c,k}] [\hat{R}_{m,k}^{-1}] = C_k [\hat{R}_{c,k} \hat{R}_{m,k}^{-1}] = C_k \hat{F}_k, \quad (5)$$

where $\hat{F}_k = \hat{R}_{c,k} \hat{R}_{m,k}^{-1}$ is the *coverage correction factor* in cell k . Factor \hat{F}_k combines the effects of overcoverage and undercoverage of cell k estimated from samples S_p and S_E . We note that undercoverage in one domain may be offset by overcoverage in the same domain. Thus, nil net undercoverage in a domain does not mean that no coverage deficiency exists in it.

The proposed estimates are based on the assumptions of the dual model, the choice of estimation cells, and the choice of the statuses defining the estimators $\hat{R}_{c,k}$ and $\hat{R}_{m,k}$. The dual model is useful since it takes into account the fact that some persons are reached neither by the census (capture) nor by the coverage survey (recapture). However, a series of conditions must be met to avoid estimation biases. The coverage survey and the census must be totally independent. The matching must be of very high quality. The model must be applied in cells with persons who have the same probability of being enumerated in the census and the survey respectively; see Section 3.3. Lastly, the population must not change too much between Census Day and the day of the survey. As to the estimators $\hat{R}_{c,k}$ and $\hat{R}_{m,k}$, they are based on the quality of the matching and the search for erroneous records. Also, it is necessary to ensure that the definition of a correct match in S_p and the definition of a correct record in S_E are identical, i.e., that there is a balance between overcoverage and undercoverage; see Section 4. All those elements are taken into consideration insofar as possible in the present estimates.

The estimate of net undercoverage in a domain d has the form $\hat{R}_{\text{netunder},d} = 1 - \hat{R}_{\text{net},d} = 1 - C_d / \hat{N}_d$, where C_d is

the enumerated number in the domain and \hat{N}_d is the estimate of the true total. The estimate of the true total \hat{N}_d is based on a *synthetic* model that assumes that the correction factor is fixed in each cell $k = 1, \dots, K$:

$$\hat{N}_d = \sum_{k=1}^K \hat{N}_{k,d} = \sum_{k=1}^K C_{k,d} \hat{F}_k. \quad (6)$$

$C_{k,d}$ is the number enumerated in the population of interest in the intersection between cell k and domain d , and \hat{F}_k is the correction factor for the coverage in cell k . The hypothesis of the synthetic model is satisfied if the behaviour of any subset in the cell is identical to that of the entire cell. This homogeneity is best controlled by the choice of cells. Here we are using the homogeneous cells defined by the dual model.

3.3 Estimation cells

The estimation cells $k = 1, \dots, K$ are constructed in such a way as to group together elements that have homogeneous probabilities of enumeration in the census and the survey respectively (dual hypothesis) and homogeneous net coverage rates (synthetic hypothesis). We want a minimum of 100 persons per cell in S_E and S_P in order to control the variance and limit the estimation bias. The variables defining the cells are selected using a logistic regression model and a discrimination method applied to the data from S_P (binary variable: P_j). The three most influential variables are cross-tabulated: nationality in two categories, marital status in two categories and size of commune in three categories. The other variables are then successively integrated. Groupings are created when the cell sizes are too small (official language of commune in two categories, age class in seven categories and sex in two categories). In the end, 121 estimation cells are obtained; see Renaud (2004) for more details.

3.4 Variance of coverage estimators

The variance of the estimators is estimated by a stratified jackknife applied to the (identical) primary units of S_P and S_E . We note that the variance of the estimated undercoverage $\hat{R}_{\text{under}} = 1 - \hat{R}_m$ is equal to the variance of the estimated matching rate \hat{R}_m . Similarly, the variance of the overcoverage $\hat{R}_{\text{over}} = 1 - \hat{R}_c$ is equal to that of the correct record rate \hat{R}_c , and the variance of the net undercoverage $\hat{R}_{\text{netunder}} = 1 - \hat{R}_{\text{net}}$ is equal to that of the net coverage rate \hat{R}_{net} .

Let θ be the parameter of interest taking the form of a weighted mean of statuses in the case of undercoverage and overcoverage, and the form of a linear function of quotients between two weighted means in the case of net undercoverage. Its estimator is $\hat{\theta}$.

Let $h = 1, \dots, H$ be the stratum used in the first stage of sampling, $i = 1, \dots, m_h$ the number of the primary unit in stratum h (postal code for NORD or commune for TICINO), and $j = 1, \dots, n_{hi}$ the number of the person in primary unit i of h . For the needs of the jackknife method, samples S_P and S_E are partitioned, in each stratum h , in m_h subsets corresponding to the persons in primary units $\alpha = 1, \dots, m_h$.

Let $\hat{\theta}_{(h\alpha)}$ be the estimator having the same form as $\hat{\theta}$ but calculated on the sample from which primary unit α of stratum h has been removed. We note that estimators $\hat{R}_{m(h\alpha),k}$ and $\hat{R}_{c(h\alpha),k}$, $k = 1, \dots, K$ are combined to form $\hat{R}_{\text{net}(h\alpha)}$:

$$\hat{R}_{\text{net}(h\alpha)} = C \left[\sum_{k=1}^K C_k \frac{\hat{R}_{c(h\alpha),k}}{\hat{R}_{m(h\alpha),k}} \right]^{-1}. \quad (7)$$

The corrected weights w'_{hij} used to calculate values $\hat{R}_{m(h\alpha)}$ and $\hat{R}_{c(h\alpha)}$ have the following form:

$$w'_{hij} = \begin{cases} 0 & \text{if } i = \alpha \\ w_{hij} \frac{m_h}{m_h - 1} & \text{if } \alpha \in h \text{ and } i \neq \alpha \\ w_{hij} & \text{if } \alpha \notin h. \end{cases} \quad (8)$$

This form of correction is preferred to the quotient between the sum of the weights of the elements in the stratum and the sum of the weights without primary unit α since it allows us to take account of the variability due to the unknown number of elements in the stratum.

The jackknife estimator becomes:

$$\hat{\theta}_{JK} = \sum_h \sum_{\alpha=1}^{m_h} \frac{\hat{\theta}_{h\alpha}}{m_h}, \quad (9)$$

with pseudo values $\hat{\theta}_{h\alpha} = m_h \hat{\theta} - (m_h - 1) \hat{\theta}_{(h\alpha)}$. The estimator of its variance can take different forms; see the example of Shao and Tu (1995). We apply the following form:

$$v(\hat{\theta}_{JK}) = \sum_h \frac{m_h - 1}{m_h} \sum_{\alpha=1}^{m_h} (\hat{\theta}_{(h\alpha)} - \hat{\theta}_{(h)})^2, \quad (10)$$

with $\hat{\theta}_{(h)} = \sum_{\alpha=1}^{m_h} \hat{\theta}_{(h\alpha)} / m_h$. Lastly, we use $v(\hat{\theta}_{JK})$ as an estimator of the variance of $\hat{\theta}$. The estimates in the subgroups use the same form of estimator with integration of a domain indicator in the construction of $\hat{\theta}_{(h\alpha)}$. No correction for the finite population is applied in the estimates. Also, other variabilities are not taken into account, such as the variability induced by the weighting model for non-response in S_P .

Problems, such as the lack of stability of estimation in strata with few primary units, appeared in the course of applying this approach. However, tests on the sharing of some primary units and a comparison with the Taylor

linearization or a simple jackknife suggest that the estimators of variance by stratified jackknife that are presented in this document are fairly conservative.

4. Choice of correct matching statuses and enumerations

A key element of coverage estimates is the definition of the *correct matching status* for the elements of S_p and the *correct enumeration status* for the elements of S_E . These correct statuses are defined on the basis of frames P_j and E_j determined during the matchings.

Is a match with a census element that is part of a collective household accepted as a correct match for an element of S_p , or is this a case of undercoverage of the population of interest? Is a duplicate outside the population of interest for an element of S_E really considered a duplicate, and hence an instance of overcoverage, or should it be excluded? A clear definition is needed. Also, the statuses used in estimates of net undercoverage must be chosen in such a way as to satisfy the balance between over- and undercoverage; see the concept of “balancing,” as, for example, in Hogan (2003). A match ($P_j = 1$) with an element outside the population of interest may, for example, be rejected as a correct match (correct match status = 0, no undercoverage) only if the search for correct records would also detect this element as incorrect because it is out of scope (correct enumeration status = 0, no overcoverage).

The criteria for defining correct statuses are constructed using information available for elements of S_p and S_E . As regards S_p , we start with the assumption that census records that were matched with elements of S_p serve to identify persons (completeness) and these persons should indeed have been enumerated (appropriateness). We also consider that they are unique, since uniqueness, while controlled by matches, is achieved in the great majority of cases controlled in S_E . The criteria of belonging to the population and correctness of location are controlled by comparison with the information collected in the coverage survey, considered as reference information. No supplementary data collection was organized to resolve ambiguous cases. As regards S_E , we have the criterion of completeness considered as having been met in the census data and the results concerning uniqueness and appropriateness obtained in the matching with the rest of the census. For duplicates and triplicates, we define $E_j = 1/d'$, with d' = number of duplicates/triplicates in the population of interest according to the census. The criteria of belonging to the population of interest and correctness of location for the elements of S_E cannot be controlled, since we have no reference data supplementing the census.

For estimates of net undercoverage, it is important to meet the balancing requirement. The criteria used in defining correct statuses are thus completeness, appropriateness and uniqueness. The criteria of belonging to the population of interest and correctness of location cannot be considered, since they are not usable in defining the correct enumeration status. The criteria of completeness, appropriateness and uniqueness are already integrated into the construction of the basic statuses P_j and E_j . Thus, estimates are made with basic statuses P_j and E_j .

For estimates not using the dual system and the need for balancing, it is possible to use other criteria to define correct statuses. Other types of correct match statuses are used in the analysis of potential measurement errors in Section 5 and the more detailed analyses of matches and enumerations presented in Renaud (2004).

5. Comparison of matches

5.1 Potential measurement errors

Measurement errors or classification errors are related to coverage errors. A person who is classified in domain d according to the census (e.g., a person between 10 and 19 years of age) but who in reality is outside the domain (e.g., a person 60 years of age) would end up as a case of overcoverage in domain d and an undercoverage case outside that domain. This misclassification does not cause a coverage error at the overall level, but it causes an error at the level of subgroups of the population.

The reasons for differences between the values collected in two surveys such as the census and the coverage survey may be quite varied and difficult to dissociate. It is to be expected that there will be matching errors, differences resulting from collection methods (paper questionnaire or telephone/face-to-face interview) and data processing methods, or real differences due to the time lag between the collection periods (December 2000 and April-May 2001). Also, it is difficult to determine the correct response if there are two different values. What is the correct choice - the census? the survey? another value not collected?

Potential measurement errors with respect to census data are analysed on the basis of a set of matches between the independent sample S_p and the census. We choose to determine which variables show respectively few or many potential classification problems, without making a judgment on the quality of either data collection. One use of this information is to evaluate the choice of estimation cells for the dual system and select subgroups for which the estimates of coverage deficiencies are soundest.

For the category variable X , we define the matching rate in the good domain \hat{R}_X as follows:

$$\hat{R}_X = \frac{\sum_{j \in S_p, \text{match}} w_{p,j} P_{X,j}}{\sum_{j \in S_p, \text{match}} w_{p,j}}, \quad (11)$$

where $w_{p,j}$ is the weight of element j of matched sample S_p (S_p, match) and the *classification status* $P_{X,j}$ is equal to 1 if element j appears in the same class in the census and the survey, and 0 otherwise. The value of \hat{R}_X is estimated with the set of matched elements and with the subgroup of elements without imputation in the census.

We also define a measure of asymmetry $\phi_X(d, d')$ for classes d and d' of variable X :

$$\phi_X(d, d') = \frac{\sum_{j \in S_p, \text{match}} w_{p,j} I_j(d, d')}{\sum_{j \in S_p, \text{match}} w_{p,j} I_j(d', d)}, \quad (12)$$

where $I_j(d, d') = 1$ if element j appears in domain d according to the survey and in domain d' according to the census, and 0 otherwise. The factor $\phi_X(d, d')$ is equal to 1 if there is a balance in the classification errors - in other words, if the number of elements in d according to the survey and d' according to the census is equal to the number in d' according to the survey and in d according to the census. The further the factor lies from 1, the less balance there is.

5.2 Potential location errors

Comparisons between the census and the survey can also be used to study people's geographic location. In the census data, we have a unique address if the person has a single residence and two addresses - principal and secondary - if the person has two residences. In the survey data, we have one or two addresses on Census Day, one or two addresses on the day of the survey and information on a possible move between the two dates. If a person has a single residence and has not moved, that person's principle address on Census Day and his/her principle address on the day of the survey are identical. The person does not have secondary addresses.

Different measures of distance are considered in order to determine potential location errors in the census. For practical reasons, including the data available, we define geographic areas around the person's principle address collected in the survey for Census Day (*reference address*). The areas are sets of political communes. They are defined on the bases of postal codes identified in the survey. The person's *basic area* is defined by the set of communes that have buildings within the postal code of the person's reference address. The definition of this area uses data from the Swiss building register, since the latter has information on buildings' postal address and the commune within which they are located. The *extended area* includes the communes within the basic area and the set of communes adjacent to them; see Renaud (2004) for examples.

Like classification errors, location errors do not cause coverage errors at the overall level but they cause errors at the level of subgroups such as regions or types of communes. Different rates may be defined. We will retain the basic location rate and the extended location rate, both weighted by $w_{p,j}$, the weight of element j of matched sample S_p . The location status takes on the value 1 if the element lies within the basic area or the extended area, as the case may be; otherwise it equals 0. In particular, we will study the correctness of location of persons who have moved, in order to detect possible problems relating to the time lag between Census Day and the actual day of collection of census data.

6. Results

6.1 Estimates of coverage deficiencies

The overall net undercoverage rate is estimated at 1.41% with a standard deviation of 0.12%. The overcoverage rate is 0.35% (standard deviation = 0.03%) and the undercoverage rate is 1.64% (standard deviation = 0.11%). These results are of the same order of magnitude as those of other countries, although they are in the lower range; see Table 1.

Overcoverage is minor in the great majority of the domains studied. The highest rate is observed for persons between 20 and 31 years of age (0.93% with a standard deviation of 0.09%); see Table 2. However, undercoverage is high in several domains. For example, a rate of 8.03% (standard deviation = 0.85%) is observed for foreigners with temporary settlement permits ("other permits") and a rate of 3.50% (standard deviation = 0.50%) is observed for 20-31-year-olds. Also, an undercoverage rate of 2.4% is observed in the Italian-speaking region of the country (language of commune: Italian; NUTS region: Ticino, and collection method: TICINO). However, the results are related to relatively great variability (standard deviation of approx. 0.5%), since samples S_p and S_E include only 1,500 and 1,700 persons respectively in this region.

Net undercoverage is positive in all the domains studied. There is therefore no net overcoverage. The highest values are observed for foreigners with permanent or temporary permits (2.89% and 3.48%, standard deviations = 0.32 % and 0.39%) as well as for 20-31-year-olds (2.84%, standard deviation = 0.36%). No significant difference is observed between males and females, between languages or between NUTS regions. Because of the small size of the sample with the collection variant TICINO, this method cannot be differentiated from the others used in the country. On the other hand, significant differences are observed between marital statuses, as well as between types and sizes of communes.

Table 1

International comparison of overall results. Estimated rates of overcoverage \hat{R}_{over} , undercoverage \hat{R}_{under} and net undercoverage $\hat{R}_{netunder}$ with corresponding estimated standard deviations. References: Statistics Canada (1999, 2004), Hogan (1993, 2003), McLennan (1997) and Trewin (2003)

		Overcoverage [%]	Undercoverage [%]	Net undercoverage [%]
Switzerland	2000	0.3 (0.03)	1.64 (0.11)	1.41 (0.12)
Canada	1996	0.74 (0.04)	3.18 (0.09)	2.45 (0.10)
	2001	0.96 (0.05)	3.95 (0.13)	2.99 (0.14)
United States	1990	3.1	4.7	1.6 (0.10)
	2000	-	-	-0.5 (0.21)
Australia	1996	0.2	1.8	1.6 (0.10)
	2001	0.9	2.7	1.8 (0.10)

We note that the net undercoverage rate is greater than the undercoverage rate in the case of permanent settlement permits. This effect, which is unrealistic, is due to the choice of estimation cells and the resulting smoothing. The construction of the cells made it necessary to group foreigners with permanent and temporary permits into a single category for aggregates so as to obtain the minimum size of 100 persons per cell. By making this grouping, we are treating foreigners as a homogeneous group, whereas this is not the case. This shows the limitations of the method and the difficulty of satisfying the assumptions of the models used in applying the approach. In the case of foreigners, we note, however, that the confidence intervals of the undercoverage and net undercoverage rates overlap. The consequences of the weaknesses of the application are therefore limited.

It should also be noted that the results are presented in domains defined by variables for which low levels of potential measurement errors were observed. The fact is that results for groups as defined by household or labour market characteristics would not be very reliable; see Section 6.2.

The precision of the results obtained is generally better than the objective set at the beginning of the project. That objective was to have a standard deviation of 0.3% for subgroups of 10,000 individuals in S_p . In the case of, for example, age classes 32-44 and 45-59, which have between 10,000 and 12,000 persons, the standard deviations are 0.19% and 0.14 %.

6.2 Potential measurement and classification errors

Of the 49,107 elements matched between the coverage survey and the census, 96% exhibit no difference in sex, the seven age classes, the three marital status classes and the three settlement permit classes (Swiss, permanent, temporary). The matching rate in the good domain \hat{R}_X is 99.3 % for sex (with and without imputations), 98.3% for marital status (98.4% for non-imputed values) and 98.7%

for settlement permits (98.8% for non-imputed values). The \hat{R}_X rate is 99.5% for age classes (with or without imputations). However, it should be noted that date of birth, along with surname and given name, was one of the main variables in the matching. Age differences are therefore possible only in the case of a non-automatic (computer-assisted or manual) matching. Three variables exhibit a matching rate in the good domain that is markedly lower than that observed for sex, age, permit and marital status. These are the variables for labour market status (in the labour market, unemployed, not in the labour market), position in household (alone, spouse, common-law union, person with child or children, other head of household, related to head of household, other; results limited to private households), and size of the person's household (according to economic residence and in private households). The \hat{R}_X rate is 90.4% for labour force status (91.1% for non-imputed values), 91.4% for position in household (94.9% for non-imputed values) and 88.3% for household size.

The measure of asymmetry $\phi_X(d, d')$ takes on the value 1.33 for sex (d = male and d' = female). There are more persons coded as males according to the survey who are coded as females in the census than there are females according to the survey who are males according to the census. The proportion of males is slightly higher in the survey. However, these results must be interpreted with caution, since they based on very few cases; see Table 3. A McNemar test is just significant at the 5% level without taking the design into account, but it is no longer significant at that level if the design is factored in. On the other hand, quite substantial asymmetries are observed for marital status. There are fewer single persons in the survey who are married in the census than the reverse (factor 0.33 for d = single and d' = married). Similarly, there are fewer married persons in the survey who are widowed in the census than the reverse (factor 0.42 for d = married and d' = other). Asymmetry is also observed for the settlement permit variable. The tendency is to have more Swiss persons in the survey who are described as foreigners in the census than the reverse, and to have more permanent permits in the survey and temporary permits in the census than the reverse (factors 5.22 for d = Swiss and d' = foreigner with permanent permit and 3.83 for d = foreigner with permanent permit and d' = foreigner with temporary permit). The factors calculated are based on few cases. However, they give an insight into the potential differences between data collection via the census questionnaire and a survey conducted mainly by telephone. The labour force status variable includes more divergent cases; see Table 4. Thus, for example, we observe fewer persons employed force in the survey and fewer persons not in the labour force census than the reverse (factor of 0.46 for d = in labour

force and d' = not in labour force). There are also fewer unemployed persons in the survey and persons not in the labour force in the census than the reverse (factor of 0.26 for d = unemployed and d' = not in labour force). The position-in-household variable also exhibits asymmetries, but these are based on few elements, since the dispersion of

the elements in the boxes (d , d') is sizable. The census variables at the household level (position in household and size of household) are influenced by the complex process of household formation. They are less reliable than those concerning persons. The values at the household level are more reliable in the survey.

Table 2 Enumerated number C and estimated rates of overcoverage \hat{R}_{over} , undercoverage \hat{R}_{under} and net undercoverage $\hat{R}_{netunder}$ for different domains [%], with corresponding estimated standard deviations (SDs)

Variable	Category	C	\hat{R}_{over}	SD	\hat{R}_{under}	SD	$\hat{R}_{netunder}$	SD
Overall		7,121,626	0.35	0.03	1.64	0.11	1.41	0.12
Sex	Male	3,497,940	0.37	0.04	1.74	0.13	1.46	0.13
	Female	3,623,686	0.33	0.03	1.55	0.10	1.37	0.13
Age class	≤ 9	810,373	0.26	0.05	1.46	0.21	1.34	0.26
	10-19	833,185	0.27	0.05	1.30	0.19	1.04	0.22
	20-31	1,115,804	0.93	0.09	3.50	0.34	2.84	0.36
	32-44	1,544,721	0.33	0.05	1.65	0.16	1.43	0.19
	45-59	1,431,771	0.22	0.04	1.18	0.14	1.04	0.14
	60-79	1,146,709	0.10	0.03	0.91	0.13	0.82	0.12
	≥ 80	239,063	0.11	0.06	1.20	0.31	1.03	0.27
Settlement permit	Swiss	5,674,266	0.33	0.03	1.28	0.09	0.98	0.10
	Foreigner, permanent	1,020,242	0.33	0.06	1.85	0.29	2.89	0.32
	Foreigner, temporary	427,118	0.56	0.11	8.03	0.85	3.48	0.39
Marital status	Single	2,975,643	0.50	0.05	2.07	0.18	1.72	0.19
	Married	3,377,223	0.23	0.04	1.27	0.11	1.25	0.12
	Widowed	369,339	0.25	0.08	1.23	0.26	0.79	0.13
	Divorced	399,421	0.24	0.08	1.95	0.35	1.02	0.10
Commune language	German + Romansh	5,128,353	0.33	0.04	1.50	0.11	1.28	0.12
	French	1,680,062	0.35	0.06	1.89	0.25	1.79	0.27
	Italian	313,211	0.53	0.12	2.35	0.49	1.56	0.19
NUTS region	Région lémanique	1,296,464	0.37	0.07	2.19	0.38	1.84	0.28
	Espace Mittelland	1,640,489	0.35	0.09	1.39	0.15	1.25	0.10
	Nordwestschweiz	976,699	0.18	0.04	1.50	0.27	1.32	0.12
	Zurich	1,221,014	0.31	0.05	1.58	0.19	1.46	0.13
	Ostschweiz	1,020,897	0.40	0.07	1.29	0.23	1.24	0.12
	Zentralschweiz	665,904	0.36	0.06	1.57	0.25	1.19	0.12
	Ticino	300,159	0.54	0.12	2.38	0.52	1.57	0.19
Commune size	Small	1,372,958	0.34	0.05	1.50	0.15	1.12	0.14
	Medium	2,398,256	0.41	0.07	1.32	0.16	1.07	0.19
	Large	3,350,412	0.31	0.03	2.01	0.19	1.77	0.19
Type	City/town	2,078,780	0.35	0.04	1.96	0.17	1.82	0.20
	Agglomeration	3,145,541	0.36	0.06	1.49	0.19	1.34	0.12
	Rural	1,897,305	0.32	0.04	1.56	0.17	1.07	0.12
Collection method	TRADITIONAL	265,607	0.39	0.05	1.91	0.28	1.07	0.12
	SEMI-TRADITIONAL	174,501	0.37	0.08	1.07	0.24	1.16	0.13
	TRANSIT + FUTURE	6,381,359	0.33	0.03	1.62	0.11	1.42	0.12
	TICINO	300,159	0.54	0.12	2.38	0.52	1.57	0.19

Table 3 Comparison of values collected in the survey and the census for the sex variable

Sex			Survey		
			Male	Female	Total
Census	Not matched	Total	393	383	776
	Matched	Total	24,171	24,936	49,107
	Matched	Male	23,967	166	24,133
		Female	204	24,770	24,974
	Matched (imputed value)	Male	6	0	6
		Female	0	13	13
Total			24,564	25,319	49,883

Table 4 Comparison of values collected in the survey and the census for the labour force status variable

Labour force status			Survey				Total
			Employed	Unemployed	Not in labour force	≤ 15 years of age	
Census	Not matched	Total	424	23	217	112	776
	Matched	Total	25,163	498	14,501	8,945	49,107
	Matched	Employed	23,953	188	2,007	13	26,161
		Unemployed	300	221	323	1	845
		Not in labour force	901	89	12,143	18	13,151
		≤ 15 years of age	9	0	28	8,913	8,950
	Matched (imputed variable)	Employed	564	22	312	6	904
		Unemployed	14	8	26	1	49
		Not in labour force	92	15	881	5	993
		≤ 15 years of age	0	0	0	0	0
Total			25,587	521	14,718	9,057	49,883

6.3 Potential location and time lag errors

Of the 49,107 elements matched between the coverage survey and the census, 97.7% are found within the basic area around the reference address collected in the survey. The corresponding value is 98.1% for persons who did not indicate any move between Census Day and the day of the survey. It is 83.9% for those who indicated a move (1,512 persons); see absolute numbers in Table 5.

It is worth noting that 9.4% of the persons in NORD who did not move are found close to their reference address but not in exactly the same building. While these problems of exact location have a negligible effect on the census data, they show the difficulty of identifying the buildings sampled when constructing lists of households in the field during the survey, as well as the difficulty of assigning persons to buildings during the processing of the census data. However, a supplementary survey would be needed to evaluate the respective effects of these two difficulties.

Efforts to locate persons who moved indicate that $151 = 145 + 6$ persons were located near their address reported on the day of the survey and not near their Census Day address (9%, weighted). Also, a set of 688 persons in NORD, among the 922 located in the two basic areas, were actually found to be residing in the building on the day of the survey. During the coverage survey, special care was taken regarding questions on addresses on Census Day and on the day of the survey. We therefore believe that the addresses of persons who moved are of better quality in the survey data than in the census data. On this basis, we deduce that out of the 1,512 persons who moved, at least $151 + 688 = 839$ are enumerated in the census at an address that they did not have on the official day of data collection but at an address that they had some time after that date. The exact time lag is not known, since the moving date was not collected in the survey.

Table 5

Comparison of the location of matched persons. The areas are defined for the address on Census Day (according to information collected in the survey) and for the address on the day of the survey (also according to information collected in the survey). Presence in the basic area, the extended area (outside the basic area) or outside the extended area for persons who did not move (stayed) and persons who moved (moved) between the census and the survey

		Day of survey				
		Stayed	Moved			Total
		Basic area	Basic area	Extended area	Outside extended area	
Census	Basic area	46,689	922	69	277	1,268
Day	Extended area	258	42	4	3	49
	Outside extended area	648	145	6	28	179
	Missing	0	15	1	0	16
	Total	47,595	1,124	80	308	1,512

7. Conclusion

Overall coverage deficiencies in the 2000 census of population in Switzerland are of the same order of magnitude as those for the censuses of other countries. However, differences are noted for subgroups (e.g., regions). Of the three components, undercoverage is of great interest, since it not only serves to detect groups of persons not well enumerated, but it also lends itself to analysing location and measurement errors. As to overcoverage estimates, these are limited by the lack of information supplementary to the census for S_E . In the future, they could be improved by collecting supplementary information on characteristics reported on Census Day in a survey of persons in that sample (e.g., location and household type). Net undercoverage estimates are based on several assumptions. The results in large domains seem reliable, but certain risks, notably related to the choice of estimation cells, exist when domains are smaller. For future estimates, we propose to evaluate the model approach applied in the United Kingdom instead of the estimation cells traditionally used in the United States.

An important element to review for future estimates is the choice of the population of interest. The decision to limit that population to persons in private households and the economic residence led to a few problems in the estimates, since it was difficult to delimit that population precisely. In a future estimation, collective households could be excluded so as to avoid practical problems relating to collection but retain all types of residences. The set of records for the economic residence would then be treated as a domain.

Estimating the coverage deficiencies of a census is an ambitious project that has proved to be worthwhile. The results provide information on the quality of the data from the 2000 census and the different coverage problems. Upcoming censuses will essentially be based on registers. Coverage estimates will be based on the experience acquired in making the 2000 estimates, with probable

adaptations to take account of the new data collection system.

Acknowledgements

I wish to thank Philippe Eichenberger of the statistical methods unit of the Office fédéral de la statistique for the productive discussions throughout the project. Warm thanks are also extended to Dr. Rajendra Singh and his colleagues in the Decennial Statistical Studies Division of the U.S. Census Bureau for their assistance in developing methods and estimates. I also wish to thank to all census personnel who performed tasks and provided information needed to carry out the project, and to Paul-André Salamin of the statistical methods unit for his careful rereading of the article.

References

- ABS (1997). The 1996 census of population and housing. Annual report 1996-97, Australian Bureau of Statistics.
- Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J. and Teague, A.D. (1999). A methodological strategy for a one-number census in the UK, *Journal of the Royal Statistical Society, Series A*, 162(2), 247-267.
- Fienberg, S.E. (1992). Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology*, 18(1), 143-154.
- Hogan, H. (1993). The post enumeration survey: Operation and results. *Journal of the American Statistical Association*, 88(423), 1047-1060.
- Hogan, H. (2003). The accuracy and coverage evaluation: Theory and design. *Survey Methodology*, 29(2), 129-138.
- McLennan, W. (1997). Census of Population and Housing, Data Quality - Undercount. Australia 1996. Information paper 2940.0. Australian Bureau of Statistics.

- Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. *Proceedings of the Survey Research Methods Section* [CD-ROM], American Statistical Association.
- Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la surcouverture (E-sample). Methodology Report 338-0019, Federal Statistical Office.
- Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Methodology Report 338-0027, Swiss Federal Statistical Office.
- Renaud, A., and Eichenberger, P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Methodology Report 338-0009, Federal Statistical Office.
- Renaud, A., and Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Methodology Report 338-0023, Federal Statistical Office.
- Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.
- Statistics Canada (1999). Couverture. Rapport technique du recensement de 1996, 92-370-XIF, Statistics Canada.
- Statistics Canada (2004). Couverture. Rapport technique du recensement de 2001, 92-394-XIF, Statistics Canada.
- Trewin, D. (2003). Census of Population and Housing, Data Quality - Undercount. Australia 2001. Information paper 2940.0. Australian Bureau of Statistics.
- Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 338-346.

Use of a web-based convenience sample to supplement a probability sample

Marc N. Elliott and Amelia Haviland¹

Abstract

In this paper we describe a methodology for combining a convenience sample with a probability sample in order to produce an estimator with a smaller mean squared error (MSE) than estimators based on only the probability sample. We then explore the properties of the resulting composite estimator, a linear combination of the convenience and probability sample estimators with weights that are a function of bias. We discuss the estimator's properties in the context of web-based convenience sampling. Our analysis demonstrates that the use of a convenience sample to supplement a probability sample for improvements in the MSE of estimation may be practical only under limited circumstances. First, the remaining bias of the estimator based on the convenience sample must be quite small, equivalent to no more than 0.1 of the outcome's population standard deviation. For a dichotomous outcome, this implies a bias of no more than five percentage points at 50 percent prevalence and no more than three percentage points at 10 percent prevalence. Second, the probability sample should contain at least 1,000-10,000 observations for adequate estimation of the bias of the convenience sample estimator. Third, it must be inexpensive and feasible to collect at least thousands (and probably tens of thousands) of web-based convenience observations. The conclusions about the limited usefulness of convenience samples with estimator bias of more than 0.1 standard deviations also apply to direct use of estimators based on that sample.

Key Words: Bias; Composite estimator; Calibration.

1. Introduction

Web-based surveys have steadily increased in use and take a variety of forms (Couper 2000). For instance, web-based probability samples use a traditional sampling frame and provide web-mode as one response option or the only response option. Web-based probability samples can have high response rates and produce estimators with minimal non-response bias (Kypri, Stephenson and Langley 2004). In contrast, web-based convenience samples are based on "inbound" hits to web pages obtained from anyone online who finds the site and chooses to participate (sometimes as a result of advertising to a population that is not specifiable) or based on volunteerism from recruited panels that are not necessarily representative of the intended population.

The primary appeal of web-based convenience samples lies in the potentially very low marginal cost per case. Visits to a web site do not require expensive labor (as for phone calls) or materials (as for mailings) for each case, combined with rapid data collection and reductions in marginal data processing costs per case. Even with some fixed costs, the total costs per case are potentially very low, especially for large surveys. The disadvantage of these samples is also clear: potentially large and unmeasured selection bias.

Most discussions of web-based convenience samples of which we are aware have either argued that probability samples are unimportant in general, tried to delineate the circumstances under which convenience samples may be useful, or dismissed the use of convenience samples entirely. We explore a different avenue by investigating the

possibility of integrating web-based convenience samples into the context of probability sampling.

In this paper we describe a methodology for combining a convenience sample with a probability sample to produce an estimator with a smaller mean squared error (MSE) than estimators that employ only the probability sample. We then explore the properties of the resulting composite estimator, a linear combination of the convenience and probability samples with weights determined by bias. This leads to recommendations regarding the usefulness of supplementing probability samples with web-based convenience samples. Because the marginal costs of web-based convenience samples are very low, we focus on identifying situations in which the increase in effective sample size (ESS) attributable to the inclusion of the convenience sample may be sufficient to justify a dual-mode approach. We demonstrate that there are limited circumstances under which a supplemental web-based convenience sample may meaningfully improve MSE. While we focus on web-based convenience samples, the discussion that follows applies to other low-cost data collection methods with poor population coverage.

2. Problem context

2.1 Initial conditions

For the combined probability/convenience sample, we propose that the same survey be administered simultaneously to a traditional probability sample (with or

1. Marc Elliott, Ph.D. and Amelia Haviland, Ph.D., Rand Corporation, 1776 Main Street, Santa Monica, CA 90407, U.S.A.

without a web-based response mode) and a web-based convenience sample. We envision a multi-purpose survey with a number of survey outcomes. In this paper we will focus on the estimation of means, but future work might extend these results to other parameters, such as regression parameters. Although we will initially consider cases where the bias of convenience sample estimates is known, we will later consider the extent to which the probability sample provides a means of measuring the unknown bias in each parameter estimate from the convenience sample.

With known bias, one may combine the convenience and probability samples in a manner that minimizes MSE. If estimates from the convenience sample are very biased, the convenience sample will accomplish little. This possibility requires that the probability sample be large enough to stand on its own. Thus, one approach would be to set aside a small portion of the probability sample budget to create a large convenience sample supplement.

For example, consider a survey for which the primary interest is in estimates for the population as a whole, but for which subpopulations estimates would also be desirable if a sample size supporting adequate precision were affordable. Suppose further that one could draw 4,000 probability observations and 10,000 convenience observations for the cost of the probability sample of 5,000. For a given outcome, if bias is large, standard errors increase moderately through a small proportionate loss in sample size; if bias is small overall and within each subpopulation, there might be a "precision windfall," allowing acceptably precise subpopulation analyses.

2.2 Initial bias reduction

We will demonstrate that the bias of convenience sample estimators must be quite small for the sample to be useful, suggesting that it may be best to focus on estimating parameters that are typically subject to less bias than overall unadjusted population estimates of proportions or means, such as regression coefficients (Kish 1985).

Additionally, one might reduce bias by calibrating the convenience sample to known population values (Kalton and Kasprzyk 1986) or by applying propensity score weights that model membership selection between the two samples to observations from the convenience sample (Rosenbaum 2002; Rosenbaum and Rubin 1983). A small set of items can be included to allow the use of either approach. These items might include both items that predict differences between respondents to web surveys and other survey modes, as well as items tailored to the content of the particular survey. The design effect from the resulting variable weights will reduce the ESS for convenience sample estimators, but the low costs of these observations makes compensating for moderate design effects affordable.

We then can estimate the remaining bias for a given parameter as the difference between the estimate in the probability sample and the weighted estimate in the convenience sample.

3. Efficiency considerations

3.1 Linear combinations of biased and unbiased estimators of a population mean

The most efficient estimator that is a linear combination of the (weighted) convenience and probability samples is a special case of an estimator given in a result by Rao (2003, pages 57-58). The properties of this estimator lead to general recommendations regarding the conditions of probability sample size, convenience sample size, and convenience sample estimator bias under which the convenience sample meaningfully improves the ESS of the probability sample.

We begin by asking: What is the most efficient estimator of this form when the magnitude of the bias is known? We will later consider relaxing the assumption of known magnitude of the bias.

Let n_1 and n_2 be the effective sample sizes of the probability sample and convenience samples, respectively, after dividing nominal sample sizes by design effects associated with the sample design and non-response adjustments. This includes propensity score or other weighting in the case of the convenience sample. The former population has mean μ , and variance σ_1^2 ; the latter has mean $\mu + \varepsilon$ and variance σ_2^2 , where ε is the known bias remaining after weighting and μ is the unknown parameter of interest. The corresponding sample means have expectation μ and $\mu + \varepsilon$ and variance σ_i^2/n_i for $i=1, 2$ under an infinite population sampling model. We assume these two estimators are uncorrelated, as they come from independent samples.

From Rao (2003, pages 57-58), the most efficient composite estimator of μ takes the form

$$\hat{\mu} = \frac{\bar{x}_2(\sigma_1^2/n_1) + \bar{x}_1(\varepsilon^2 + \sigma_2^2/n_2)}{\varepsilon^2 + \sigma_1^2/n_1 + \sigma_2^2/n_2},$$

with remaining bias

$$\varepsilon_c = \varepsilon \left(\frac{\sigma_1^2/n_1}{\varepsilon^2 + \sigma_1^2/n_1 + \sigma_2^2/n_2} \right)$$

and

$$\text{MSE}_c = \frac{(\sigma_1^2/n_1)(\varepsilon^2 + \sigma_2^2/n_2)}{\varepsilon^2 + \sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

As can be seen, the composite estimator is a convex combination of the convenience sample and probability sample means. The influence of the former is determined by the ratio of the MSE (here variance) of the probability sample mean to the sum of that term and the MSE of the convenience sample mean. Similarly, the remaining bias is the original bias multiplied by this same ratio, whereas the resultant MSE_c is the product of the two MSEs divided by their sum. Note that bias approaches zero both as $\varepsilon \rightarrow 0$ (no selection bias in the convenience sample estimate) and as $\varepsilon \rightarrow \infty$ (no weight given to the convenience sample).

3.2 Quantifying the contributions of the convenience sample

We now can evaluate the contributions of the convenience sample based on the known remaining bias in its associated estimators. To this end, we will define several quantities.

Let

$$ESS_1 = \frac{\sigma_1^2/n_1}{MSE_c} n_1 = \left(\frac{\varepsilon^2 + \sigma_1^2/n_1 + \sigma_2^2/n_2}{\varepsilon^2 + \sigma_2^2/n_2} \right) n_1$$

be the effective sample size needed for an unbiased sample mean with the same MSE as the composite estimator. To further simplify this expression, let us define the remaining standardized bias, $E = \varepsilon/\sigma$, and consider the case in which the observations from the convenience and probability populations have equal variance, $(\sigma_1 = \sigma_2 = \sigma)$. In this case, the increment to ESS_1 attributable to the convenience sample, the difference between ESS_1 with and without the convenience sample, is

$$\frac{1}{1/n_2 + E^2} = n_2 \left(\frac{1}{1 + n_2 E^2} \right).$$

3.3 Maximum contribution of the convenience sample

As $n_2 \rightarrow \infty$, the increment to ESS_1 approaches $1/E^2$. This limit, the inverse of the squared standardized bias, is the maximum possible incremental contribution of the convenience sample to the ESS_1 (abbreviated MICCS). If the MICCS is small, then a convenience sample of any size cannot meaningfully improve MSE. If the MICCS is large enough to be meaningful, we then need to consider what convenience sample sizes are needed to achieve a large proportion of the MICCS.

To develop intuition for the magnitude of E (standardized bias) we consider the important case of a dichotomous outcome, for which $E = \varepsilon/\sqrt{P(1-P)}$ where P is the population probability of the outcome. Table 1 below translates bias for a dichotomous outcome from percentage points to standardized bias and then to the corresponding MICCS for $P = 0.1$ and $P = 0.5$.

Table 1 Maximum contributions of convenience samples to the estimation of a proportion by bias in percentage points

<i>E</i> (Standardized Bias*)	Overall Prevalence of Outcome		MICCS ^{&}
	10%	50%	
	Bias in Percentage Points		
0.01	0.3%	0.5%	10,000
0.02	0.6%	1.0%	2,500
0.05	1.5%	2.5%	400
0.10	3.0%	5.0%	100
0.20	6.0%	10.0%	25

[#] Of estimators of means using only the convenience sample
[&] ESS added with an infinitely large convenience sample relative to no use of a convenience sample.

For a proportion near 50%, a bias of 2.5 percentage points limits the potential increment of ESS_1 to 400. The minimum increment to ESS_1 that offsets the fixed cost of setting up the web-based response mode will vary by user, but we suspect increments of less than 100 will rarely be cost-effective. Table 1 then implies that convenience samples for which the standardized biases of estimators restricted to the convenience sample generally exceed 0.1 standard deviations will rarely prove cost-effective. For a dichotomous variable with P between 0.1 and 0.5 this corresponds to a bias of 3 to 5 percentage points.

How easily are biases of this magnitude achieved with adjusted estimates from convenience samples? Several studies compared propensity-weighted web-based convenience samples to RDD surveys. One (Taylor 2000) advocated the stand-alone use of such convenience samples despite differences of as much as five percentage points in a number of estimates for dichotomous outcomes regarding political attitudes, with standardized bias of 0.05 to 0.10 if one treats RDD as a gold standard. Another (Schonlau, Zapert, Simon, Sanstad, Marcus, Adams, Spranca, Kan, Turner and Berry 2003) does not report magnitudes of differences, but does report that 29 of 37 items regarding health concerns exhibit differences that are statistically significant at $p < 0.01$. Given the reported sample sizes (and optimistically ignoring any DEFF from weighting), it can be shown that significance at that threshold implies point estimates of standardized bias exceeding 0.05 for estimators of 78% of items. The key outcome in a Slovenian comparison of a probability phone sample and a Web-based convenience sample (Vehovar, Manfreda and Batagelj 1999) would be estimated with a standardized bias of more than 0.1 from the convenience sample even after extensive weighting adjustments. It should be noted that there may also be mode effects on responses for the Web mode when compared to a telephone mode among subjects randomized to response mode (Fricker, Galesic, Tourangeau and Yan 2005), so that not all differences between Web convenience samples and non-Web probability samples may result from selection.

3.4 Actual contribution of the convenience sample

While the maximum possible increment (MICCS) is $1/E^2$, the actual increment to ESS_1 can be expressed as $(k/k+1)$ MICCS where $k = n_2E^2$. The shortfall of the actual increment to ESS_1 from the MICCS can then be expressed as $MICCS - ESS_1 = 1/[(E^2)(1 + n_2E^2)]$. This implies that the returns to ESS_1 diminish with increasing size of the convenience sample, more quickly with large bias since the bias eventually dominates any further variance reduction. Half of the MICCS noted is achieved when the ESS of the convenience sample is equal to the MICCS. For example, if bias is 0.01 standard deviations and a convenience sample has an ESS of 10,000, then the MICCS is 10,000, but the actual incremental contribution to ESS_1 will be 5,000. This suggests that convenience samples with ESS 2-20 times as large as MICCS will suffice for most purposes, which correspond to 67%-95% of the potential gain in ESS. Such heuristics in turn imply collecting 200 - 4,000 such cases when E is relatively large ($E = 0.05$ to 0.10) and 5,000 - 200,000 such cases when E is relatively small ($E = 0.01$ to 0.02). Table 2 provides illustrative examples of the ESS_1 achieved at several combinations of sample sizes and bias.

Table 2 Examples of ESS_1 at several sample sizes and levels of standardized bias

n_1 (Probability Sample Size)	n_2 (Convenience Sample Size)	E (Standardized Bias [#])	ESS_1 for the Composite Estimate	$ESS_1 / n_1^{\&}$
1,000	1,000	0.01	1,909	1.909
1,000	1,000	0.10	1,091	1.091
1,000	10,000	0.01	6,000	6.000
1,000	10,000	0.10	1,099	1.099
1,000	100,000	0.01	10,091	10.091
1,000	100,000	0.10	1,100	1.100
10,000	1,000	0.01	10,909	1.091
10,000	1,000	0.10	10,091	1.009
10,000	10,000	0.01	15,000	1.500
10,000	10,000	0.10	10,099	1.010
10,000	100,000	0.01	19,091	1.909
10,000	100,000	0.10	10,100	1.010

Number of estimators of means using only the convenience sample

[#] Of estimators of means using only the convenience sample

[&] ESS relative to no use of a convenience sample.

3.5 Precision for estimating bias

Heretofore, we have assumed a known bias in convenience sample estimators; in practice, the bias will need to be estimated using information from both samples. We next explore the extent to which the size of the probability sample also constrains the usefulness of the convenience sample through the need to precisely estimate the remaining bias.

We can estimate ε as the difference between the sample mean of the probability sample and the weighted mean of the convenience sample. The true standard error for the

estimate of bias is $\sigma_{\varepsilon} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. If $\sigma_1 = \sigma_2 = \sigma$, the true standard error for the estimate of standardized bias (E) is $\sigma_{\hat{E}} = \sqrt{1/n_1 + 1/n_2}$. No matter how large the convenience sample, this term can never be less than the inverse of the square root of the probability sample size.

It has been demonstrated that the relative error in MSE for a composite estimator is relatively insensitive to small errors in the estimates of bias (Schaible 1978), which is encouraging for well-estimated biases. Unfortunately, unless both the probability and convenience ESS are large, the standard error of the estimate for E is impractically large relative to the values of E that make the convenience supplement useful ($E < 0.10$). For example, suppose that a probability sample of ESS 1,000 and a convenience sample of ESS 5,000 yielded a point estimate of standardized bias of 0.02. If the point estimate were correct, the convenience sample would increase the ESS_1 by 1,667. But this estimate could also have a true bias of 0.088 standard deviations (95% upper confidence limit), which would imply that the increment would be less than 130.

If we assume that the convenience sample size will always be at least twice the probability sample size, these results imply that practical applications of this technique must have a minimum sample size of 1,000-10,000 for the probability sample if they are to address the uncertainty in the magnitude of bias in convenience sample estimators (standard errors of E in the 0.01 to 0.04 range).

4. Discussion

We describe a composite estimator that is a linear combination of an unbiased sample mean estimate from a probability sample and a biased (propensity-score weighted) sample mean estimate from a web-based convenience sample. We use the MSE of this composite estimator to characterize the contributions of the convenience sample to an estimator based only on the probability sample in terms of ESS. We then calculate the maximal contribution of the convenience sample, the role of the convenience sample size in approaching this limit, and the roles of both sample sizes in estimating bias with sufficient precision.

Practitioners sometimes assume that small probability samples are sufficient to estimate the bias in estimates from corresponding convenience samples. Our results suggest otherwise. We demonstrate that the standardized bias of web-based convenience sample estimators after initial adjustments to reduce bias must be quite small (no more than 0.1 standard deviations, and probably less than 0.05 standard deviations) for the MSE of the overall estimate to be meaningfully smaller than it would be without use of the convenience sample. We further demonstrate that convenience sample sizes of thousands or tens of thousands

are also needed to realize practical gains. Finally, we demonstrate that a large probability sample size (1,000-10,000) is also needed for reasonably precise estimates of the remaining bias in initially bias-adjusted convenience sample estimators. Because the bias of estimates in an application to a multipurpose survey is likely to vary by outcome, the global decision to substitute a large number of inexpensive surveys for fewer traditional surveys must be made carefully.

The greatest opportunity in cost savings may be in large surveys, simply as a function of their size. On the other hand, the greatest proportionate gains in precision are likely to occur for samples of intermediate size. Gains might also be substantial for large samples in which the main inferences are smaller subgroups. For example, a national survey of 100,000 individuals might make inference to 200 geographic subregions, with samples of 500 for each. If one supplemented this national sample with a very large web-based convenience sample, estimated the bias nationally, and elected to assume that the bias did not vary regionally, one might decrease the MSE of the sub-region estimates substantially through the use of such a composite estimator.

As a final caveat, the conclusions about the limited usefulness of convenience samples with estimator bias of more than 0.1 standard deviations are not limited to attempts to use a composite estimator. The same approach can be applied to show that an estimator based only on a convenience sample of any size with a standardized bias of 0.2 (e.g., ten percentage points for a dichotomous variable with $P = 0.5$) will have an MSE greater than or equal to that of an estimate from a probability sample of size 25.

Acknowledgements

The authors would like to thank John Adams, Ph.D., for his comments on the manuscript, Matthias Schonlau, Ph.D., for his comments on an earlier version of the manuscript, and Colleen Carey, B.A., Kate Sommers-Dawes, B.A., and

Michelle Platt, B.A. for their assistance in the preparation of the manuscript.

Marc Elliott is supported in part by the Centers for Disease Control and Prevention (CDC U48/DP000056); the views expressed in this article do not necessarily reflect the views of the Centers for Disease Control and Prevention.

References

- Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.
- Fricker, S., Galesic, M., Tourangeau R. and Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69, 370-392.
- Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- Kish, L. (1985). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kypri, K., Stephenson, S. and Langley, J. (2004). Assessment of nonresponse bias in an internet survey of alcohol use. *Alcoholism: Clinical and Experimental Research*, 28, 630-634.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rosenbaum, P.R. (2002). *Observational Studies*, 2nd Edition, New York: Springer-Verlag.
- Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika*, 70, 41-55.
- Schaible, W.A. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 741-746.
- Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. and Berry, S. (2003). A comparison between responses from propensity-weighted web survey and an identical RDD survey. *Science Computer Review*, 21, 1-11.
- Taylor, H. (2000). Does internet research 'work'? Comparing on-line survey results with telephone surveys. *International Journal of Market Research*, 42, 41-63.
- Vehovar, V., Manfreda, K.L. and Batagelj, Z. (1999). Web surveys: Can weighting solve the problem? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 962-967.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2007.

- R.C. Bailey, *George Mason University*
M. Bankier, *Statistics Canada*
J.-F. Beaumont, *Statistics Canada*
Y. Bélanger, *Statistics Canada*
W. Bell, *U.S. Bureau of the Census*
Y. Berger, *University of Southampton*
C. Boudreau, *Medical College of Wisconsin*
M. Brick, *Westat, Inc.*
J. Chen, *University of Waterloo, Canada*
R. Clark, *University of Wollongong, Australia*
E. Dagum, *University of Bologna*
A. Dessertaine, *EDF, R&D-OSIRIS-CLAMART*
F. Dupont, *INSEE*
M. Elliott, *University of Michigan*
J. Eltinge, *U.S. Bureau of Labor Statistics*
L.R. Ernst, *Bureau of Labour Statistics, U.S.A.*
L. Fattorini, *Università di Siena*
R. Fisher, *U.S. Department of Treasury*
W. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
D. Garriquet, *Statistics Canada*
C. Girard, *Statistics Canada*
E. Gois, *Instituto Nacional de Estatística, Portugal*
B. Graubard, *National Cancer Institute*
R. Griffin, *US Bureau of the Census*
D. Haziza, *Statistics Canada*
S. Heeringa, *ISR, Michigan*
M. Hidirolou, *Statistics Canada*
J. Jiang, *University of California at Davis*
W. Kalsbeek, *U. of North Carolina*
J.-K. Kim, *Department of Applied Statistics, Seoul, Korea*
P. Knottnerus, *Statistics Netherlands*
E. Korn, *National Institute of Health*
P. Kott, *National Agricultural Statistics Service*
D. Ladiray, *INSEE*
P. Lahiri, *JPSM, University of Maryland*
N. Laniel, *Statistics Canada*
M. Latouche, *Statistics Canada*
P. Lavallée, *Statistics Canada*
P. Lavaq, *Statistics Canada*
S. Lee, *University of California at Los Angeles*
S. Lele, *University of Alberta*
C. Léon, *Statistics Canada*
P.S. Lévy, *RTI International*
S. Lohr, *Arizona State University*
W.W. Lu, *Acadia University*
L. Mach, *Statistics Canada*
T. Maiti, *Iowa State University*
D. Malec, *United States Bureau of the Census*
H. Mantel, *Statistics Canada*
A. Matei, *Université de Neuchâtel, Switzerland*
S. Merad, *Office for National Statistics, UK*
J. Montequila, *Westat, Inc.*
R. Munnich, *University of Tubingen*
P.L. Nascimento Silva, *University of Southampton*
G. Nathan, *Hebrew University of Jerusalem*
J. Opsomer, *Iowa State University*
S.M. Paddock, *Rand corporation, U.S.A.*
S. Pramanik, *JPSM, University of Maryland*
N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel, Switzerland*
T.J. Rao, *ISI*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
S. Rubin-Bleuer, *Statistics Canada*
M. del Mar Rueda, *University of Granada*
P. Saavedra, *ORC Macro, U.S.A.*
F. Scheuren, *NORC*
E. Schindler, *US Bureau of the Census*
K. Sinha, *University of Illinois at Chicago, U.S.A.*
C.J. Skinner, *University of Southampton*
P.J. Smith, *Center for Disease Control and Prevention*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
F. Verret, *Statistics Canada*
S.G. Walker, *University of Kent, U.K.*
W. Winkler, *US Bureau of the Census*
K. Wolter, *NORC*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*

Acknowledgements are also due to those who assisted during the production of the 2007 issues: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Micheal Pelchat and Isabelle Poliquin (Dissemination Division), Sheri Buck (Systems Development Division), François Beaudin (Official Languages and Translation Division) and Sophie Chartier and Gayle Keely (Household Survey Methods Division). Finally we wish to acknowledge Christine Cousineau, Céline Ethier and Denis Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 23, No. 1, 2007

Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965 Margo Anderson and William Seltzer	1
Efficient Stratification Based on Nonparametric Regression Methods Enrico Fabrizi and Carlo Trivisano	35
A Selection Strategy for Weighting Variables Under a Not-Missing-at-Random Assumption Barry Schouten	51
Imputing for Late Reporting in the U.S. Current Employment Statistics Survey Kennon R. Copeland and Richard Valliant	69
Incentives in Random Digit Dial Telephone Surveys: A Replication and Extension Richard Curtin, Eleanor Singer and Stanley Presser	91
Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience Peter Lynn, Sabine Häder, Siegfried Gabler and Seppo Laaksonen	107
Book and Software Reviews	125

Contents

Volume 23, No. 2, 2007

Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents Jaak Billiet, Michel Phillippens, Rory Fitzgerald and Ineke Stoop	135
Measuring Disability in Surveys: Consistency over Time and Across Respondents Sunghee Lee, Nancy A. Mathiowetz and Roger Tourangeau	163
Quantifying Stability and Change in Ethnic Group Ludi Simpson and Bola Akinwale	185
Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production William P. Cleveland and Stuart Scott	209
Finite Population Small Area Interval Estimation Li-Chun Zhang	223
Predicting Natural Gas Production in Texas: An Application of Nonparametric Reporting Lag Distribution Estimation Crystal D. Linkletter and Randy R. Sitter	239
Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update Laura Zayatz	253
Book and Software Reviews	267

Contents
Volume 23, No. 3, 2007

The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science	
Michael F. Goodchild.....	269
Using Geospatial Information Resources in Sample Surveys	
Sarah M. Nusser	285
Discussion	
Linda Williams Pickle.....	291
Optimizing the Use of Microdata: An Overview of the Issues	
Julia Lane	299
Benchmarking the Effect of Cell Adjustment on Tabular Outputs: The Shortcomings of Current Approaches	
Paul Williamson	319
Summary of Accuracy and Coverage Evaluation for the U.S. Census 2000	
Mary H. Mulry.....	345
Resampling Variance Estimation in Surveys with Missing Data	
A.C. Davison and S. Sardy	371
Nonresponse Among Ethnic Minorities: A Multivariate Analysis	
Remco Feskens, Joop Hox, Gerty Lensvelt-Mulders and Hans Schmeets	387
Procedures for Updating Classification Systems: A Study of Biotechnology and the Standard Occupational Classification System	
Neil Malhotra and Jon A. Krosnick	409

All inquiries about submissions and subscriptions should be directed to jos@scb.se

Volume 35, No. 1, March/mars 2007

Paul GUSTAFSON	
Editorial	1
Report from the previous Editor	5
Acknowledgement of referees' services	7
Renato ASSUNÇÃO, Andréa TAVARES, Thais CORREA & Martin KULLDORFF	
Space-time cluster identification in point processes	9
Elvan CEYHAN, Carey E. PRIEBE & David J. MARCHETTE	
A new family of random graphs for testing spatial segregation	27
Aurélien LABBE & Mary E. THOMPSON	
Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies	51
H.a.o.l.a.n. LU, James S. HODGES & Bradley P. CARLIN	
Measuring the complexity of generalized linear hierarchical models	69
Aurore DELAIGLE	
Nonparametric density estimation from data with a mixture of Berkson and classical errors	89
Ozlem ILK & Michael J. DANIELS	
Marginalized transition random effect models for multivariate longitudinal binary data	105
Richard A. LOCKHART, John J. SPINELLI & Michael A. STEPHENS	
Cramér-von Mises statistics for discrete distributions with unknown parameters	125
P.W. FONG, W.K. LI, C.W. YAU & C.S. WONG	
On a mixture vector autoregressive model	135
Jung Wook PARK, Marc G. GENTON & Sujit K. GHOSH	
Censored time series analysis with autoregressive moving average models	151
Abdessamad SAIDI	
Consistent testing for non-correlation of two cointegrated ARMA time series	169
Forthcoming papers/Articles à paraître	189
Complete online access to <i>The Canadian Journal of Statistics</i>	191

Volume 35, No. 2, June/juin 2007

Pierre DUCHESNE	
On consistent testing for serial correlation in seasonal time series models	193
Jaakko NEVALAINEN, Denis LAROCQUE & Hannu OJA	
On the multivariate spatial median for clustered data	215
Liqun WANG	
A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models	233
Juan Carlos PARDO-FERNÁNDEZ, Ingrid VAN KEILEGOM & Wenceslao GONZÁLEZ-MANTEIGA	
Goodness-of-fit tests for parametric models in censored regression	249
Song Xi CHEN & Tzee-Ming HUANG	
Nonparametric estimation of copula functions for dependence modelling	265
Victor DE OLIVEIRA	
Objective Bayesian analysis of spatial data with measurement error	283
Gonzalo GARCÍA-DONATO & Dongchu SUN	
Objective priors for hypothesis testing in one-way random effects models	303
Hui LI, Guosheng YIN & Yong ZHOU	
Local likelihood with time-varying additive hazards model	321
Forthcoming papers/Articles à paraître	338
Complete online access to <i>The Canadian Journal of Statistics</i>	339

Volume 35, No. 3, September/septembre 2007

David F. ANDREWS Robust likelihood inference for public policy.....	341
Zeny Z. FENG, Jiahua CHEN, Mary E. THOMPSON Asymptotic properties of likelihood ratio test statistics in affected-sib-pair analysis	351
Hwashin H. SHIN, Glen TAKAHARA & Duncan J. MURDOCH Optimal designs for calibration of orientations.....	365
Jiancheng JIANG, Haibo ZHOU, Xuejun JIANG & Jianan PENG Generalized likelihood ratio tests for the structure of semiparametric additive models.....	381
Runze LI & Lei NIE A new estimation procedure for a partially nonlinear model via a mixed-effects approach.....	399
Axel MUNK, Matthias MIELKE, Guido SKIPKA & Gudrun FREITAG Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics.....	413
Inkyung JUNG & Martin KULLDORFF Theoretical properties of tests for spatial clustering of count data.....	433
QiQi LU & Robert B. LUND Simple linear regression with multiple level shifts.....	447
Forthcoming papers/Articles à paraître.....	459
Volume 36 (2008): Subscription rates/Frais d'abonnement	460

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω ; o, O, 0; l, I).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

1.

Présentation
- 1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3

Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4

Les remerciements doivent paraître à la fin du texte.
- 1.5

Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.

Résumé
- Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.
3.

Rédaction
- 3.1

Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) *etc.*
- 3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5

Distinguer clairement les caractères ambigus (comme w, ω ; o, O, 0; 1, I).
- 3.6

Les caractères italiques sont utilisés pour faire ressortir des mots.
4.

Figures et tableaux
- 4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).
5.

Bibliographie
- 5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
- 5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, *etc.* à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
6.

Communications brèves
- 6.1

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

David F. ANDREWS	Robust likelihood inference for public policy.....	341
Zeny Z. FENG, Jiahua CHEN, Mary E. THOMPSON	Asymptotic properties of likelihood ratio test statistics in affected-sib-pair analysis.....	351
Hwashin H. SHIN, Glen TAKAHARA & Duncan J. MURDOCH	Optimal designs for calibration of orientations.....	365
Jiancheng JIANG, Haibo ZHOU, Xuejun JIANG & Jiaman PENG	Generalized likelihood ratio tests for the structure of semiparametric additive models.....	381
Runze LI & Lei NIE	A new estimation procedure for a partially nonlinear model via a mixed-effects approach.....	399
Axel MUNKE, Mathias MIELKE, Guido SKIPKA & Gudrun FREITAG	Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics.....	413
Inkyung JUNG & Martin KULLDORFF	Theoretical properties of tests for spatial clustering of count data.....	433
Qi LI & Robert B. LUND	Simple linear regression with multiple level shifts.....	447
	Forthcoming papers/Articles à paraître.....	459
	Subscription rates/Frais d'abonnement.....	460

CONTENTS

TABLE DES MATIÈRES

Volume 35, No. 1, March/mars 2007

Paul GUSTAFSON

Editorial

Report from the previous Editor

Acknowledgement of referees' services

Renato ASSUNÇÃO, Andréa TAVARES, Thaís CORREA & Martin KULLDORFF

Space-time cluster identification in point processes

Elyan CEYHAN, Carey E. PRIEBE & David J. MARCHETT

A new family of random graphs for testing spatial segregation

Aurélie LABBE & Mary E. THOMPSON

Multiple testing using the posterior probabilities of directional alternatives, with application to genomic studies

H.a.o.lan, LU, James S. HODGES & Bradley P. CARLIN

Measuring the complexity of generalized linear hierarchical models

Aurore DELAIGLE

Nonparametric density estimation from data with a mixture of Berkson and classical errors

Ozlem TLK & Michael J. DANIELS

Marginalized transition random effect models for multivariate longitudinal binary data

Richard A. LOCKHART, John T. SPINELLI & Michael A. STEPHENS

Cramér-von Mises statistics for discrete distributions with unknown parameters

P. W. FONG, W. K. LI, C. W. YAU & C. S. WONG

On a mixture vector autoregressive model

Jung Wook PARK, Marc G. GENTON & Sujit K. GHOSH

Censored time series analysis with autoregressive moving average models

Abdessamad SAIDI

Consistent testing for non-correlation of two cointegrated ARMA time series

Forthcoming papers/Articles à paraître

Complete online access to *The Canadian Journal of Statistics*

Volume 35, No. 2, June/juin 2007

Pierre DUCHESNE

On consistent testing for serial correlation in seasonal time series models

Jaakko NEVALAINEN, Denis LAROCQUE & Hannu OJA

On the multivariate spatial median for clustered data

Liquan WANG

A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models

Juan Carlos PARDO-FERNÁNDEZ, Ingrid VAN KEILEGOM & Wenceslao GONZÁLEZ-MANTEIGA

Goodness-of-fit tests for parametric models in censored regression

Song Xi CHEN & Tzee-Ming HUANG

Nonparametric estimation of copula functions for dependence modelling

Victor DE OLIVEIRA

Objective Bayesian analysis of spatial data with measurement error

Gonzalo GARCÍA-DONATO & Dongchu SUN

Objective priors for hypothesis testing in one-way random effects models

Hui LI, Guosheng YIN & Yong ZHOU

Local likelihood with time-varying additive hazards model

Forthcoming papers/Articles à paraître

Complete online access to *The Canadian Journal of Statistics*

Contents

Volume 23, No. 3, 2007

269	The Morris Hansen Lecture 2006 Statistical Perspectives on Spatial Social Science Michael F. Goodchild.....
285	Using Geospatial Information Resources in Sample Surveys Sarah M. Nusser
291	Discussion Linda Williams Pickle.....
299	Optimizing the Use of Microdata: An Overview of the Issues Julia Lane.....
319	Benchmarking the Effect of Cell Adjustment on Tabular Outputs: The Shortcomings of Current Approaches Paul Williamson
345	Summary of Accuracy and Coverage Evaluation for the U.S. Census 2000 Mary H. Mulry.....
371	Resampling Variance Estimation in Surveys with Missing Data A.C. Davison and S. Sardy.....
387	Nonresponse Among Ethnic Minorities: A Multivariate Analysis Remco Feskens, Joop Hox, Gerty Lensvelt-Mulders and Hans Schmets
409	Procedures for Updating Classification Systems: A Study of Biotechnology and the Standard Occupational Classification System Neil Malhotra and Jon A. Kroznick

All inquiries about submissions and subscriptions should be directed to jos@scb.se

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

JOURNAL OF OFFICIAL STATISTICS
An International Review Published by Statistics Sweden

Contents
Volume 23, No. 1, 2007

Challenges to the Confidentiality of U.S. Federal Statistics, 1910-1965	1
Margo Anderson and William Selzer	
Efficient Stratification Based on Nonparametric Regression Methods	
Enrico Fabrizi and Carlo Trivisano	35
A Selection Strategy for Weighing Variables Under a Not-Missing-at-Random Assumption	
Barry Schouten	51
Imputing for Late Reporting in the U.S. Current Employment Statistics Survey	
Kennon R. Copeland and Richard Valliant	69
Incentives in Random Digit Dial Telephone Surveys: A Replication and Extension	
Richard Curtin, Eleanor Singer and Stanley Presser	91
Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience	
Peter Lynn, Sabine Häder, Siegfried Gabler and Seppo Laaksonen	107
Book and Software Reviews	125

Contents
Volume 23, No. 2, 2007

Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents	135
Jaak Biller, Michel Philippens, Rory Fitzgerald and Ineke Stoop	
Measuring Disability in Surveys: Consistency over Time and Across Respondents	
Sunghee Lee, Nancy A. Mathiowetz and Roger Tourangeau	163
Quantifying Stability and Change in Ethnic Group	
Ludi Simpson and Bola Akinwale	185
Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production	
William P. Cleveland and Stuart Scott	209
Finite Population Small Area Interval Estimation	
Li-Chun Zhang	223
Predicting Natural Gas Production in Texas: An Application of Nonparametric Reporting Lag Distribution Estimation	
Crystal D. Linkletter and Randy R. Sitter	239
Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update	
Laura Zayatz	253
Book and Software Reviews	267

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2007.

- R.C. Bailey, *George Mason University*
M. Bankier, *Statistique Canada*
J.-F. Beaumont, *Statistique Canada*
Y. Bélanger, *Statistique Canada*
W. Bell, *U.S. Bureau of the Census*
Y. Berger, *University of Southampton*
C. Boudreau, *Medical College of Wisconsin*
M. Brick, *Westat, Inc.*
J. Chen, *University of Waterloo, Canada*
R. Clark, *University of Wollongong, Australie*
E. Dagum, *University of Bologna*
A. Dessertaine, *EDF, R&D-OSIRIS-CLAMART*
F. Dupont, *INSEE*
M. Elliott, *University of Michigan*
J. Eltinge, *U.S. Bureau of Labor Statistics*
L.R. Erms, *Bureau of Labour Statistics, E.-U.*
L. Fattorini, *Università di Siena*
R. Fisher, *U.S. Department of Treasury*
W. Fuller, *Iowa State University*
J. Gambino, *Statistique Canada*
D. Garrigue, *Statistique Canada*
C. Girard, *Statistique Canada*
E. Gois, *Instituto Nacional de Estatística, Portugal*
B. Graubard, *National Cancer Institute*
R. Griffin, *US Bureau of the Census*
D. Haziza, *Statistique Canada*
S. Heeringa, *ISR, Michigan*
M. Hidiroglou, *Statistique Canada*
J. Jiang, *University of California at Davis*
W. Kalsbeek, *U. of North Carolina*
J.-K. Kim, *Department of Applied Statistics, Seoul, Corée*
P. Knothnerus, *Statistics Netherlands*
E. Korn, *National Institute of Health*
P. Kott, *National Agricultural Statistics Service*
D. Laditray, *INSEE*
P. Lahiri, *JPSM, University of Maryland*
N. Lange, *Statistique Canada*
M. Latouche, *Statistique Canada*
P. Lavallée, *Statistique Canada*
P. Lavaq, *Statistique Canada*
S. Lee, *University of California at Los Angeles*
S. Lele, *University of Alberta*
C. Léon, *Statistique Canada*
P.S. Lévy, *RTI International*
S. Lohr, *Arizona State University*
W.W. Lu, *Acadia University*
L. Mach, *Statistique Canada*
T. Mati, *Iowa State University*
D. Malec, *United States Bureau of the Census*
H. Mamel, *Statistique Canada*
A. Matei, *Université de Neuchâtel, Suisse*
S. Merad, *Office for National Statistics, R.-U.*
J. Montequila, *Westat, Inc.*
R. Munnich, *University of Tübingen*
P.L. Nascimento Silva, *University of Southampton*
G. Nathan, *Hebrew University of Jerusalem*
J. Opsomer, *Iowa State University*
S.M. Paddock, *Rand corporation, E.-U.*
S. Pramanik, *JPSM, University of Maryland*
N. Prasad, *University of Alberta*
L. Qualité, *Université de Neuchâtel, Suisse*
T.J. Rao, *ISI*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
M. Rubin-Bleuer, *Statistique Canada*
M. del Mar Rueda, *University of Granada*
P. Saavedra, *ORC Macro, E.-U.*
F. Scheuren, *NORC*
K. Schindler, *US Bureau of the Census*
E. Sinha, *University of Illinois at Chicago, E.-U.*
C.J. Skinner, *University of Southampton*
P.J. Smith, *Center for Disease Control and Prevention*
M. Thompson, *University of Waterloo*
Y. Tillé, *Université de Neuchâtel*
F. Verret, *Statistique Canada*
S.G. Walker, *University of Kent, R.-U.*
W. Winkler, *US Bureau of the Census*
K. Wolter, *NORC*
C. Wu, *University of Waterloo*
W. Yung, *Statistique Canada*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2007: Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Michael Pelchat et Isabelle Poliquin (Division de la diffusion), Sheri Buck (Division du développement de systèmes), François Beaudin (Division des langues officielles et traduction) et Sophie Charrier et Gayle Keely (Division des méthodes d'enquêtes auprès des ménages). Finalement nous désirons exprimer notre reconnaissance à Christine Cousineau, Céline Ethier et Denis Lemire de la Division des méthodes d'enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

Bibliographie

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Fritcher, S., Galesic, M., Tourangeau R. et Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69, 370-392.

Katlon, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Kish, L. (1985). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kyprk, K., Stephenson, S. et Langley, J. (2004). Assessment of nonresponse bias in an internet survey of alcohol use. *Alcoholism: Clinical and Experimental Research*, 28, 630-634.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Rosenbaum, P.R. (2002). *Observational Studies*, 2^{ème} édition, New York : Springer-Verlag.

Rosenbaum, P.R., et Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Schabale, W.A. (1978). Choosing weights for composite estimators for small area statistics. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 741-746.

Schonlau, M., Zapert, K., Simon, L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R. et Berry, S. (2003). A comparison between responses from propensity-weighted web survey and an identical RPD survey. *Science Computer Review*, 21, 1-11.

Taylor, H. (2000). Does internet research 'work'? Comparing on-line survey results with telephone surveys. *International Journal of Market Research*, 42, 41-63.

Vehovar, V., Manfeda, K.L. et Batagelj, Z. (1999). Web surveys: Can weighting solve the problem? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 962-967.

Mentionnons maintenant une dernière mise en garde. Les conclusions au sujet de l'utilité limitée des échantillons de convenance lorsque le biais des estimateurs est supérieur à 0,1 fois les écarts-types ne se limitent pas aux tentatives visant à utiliser un estimateur composite. La même approche peut être appliquée pour démontrer que, dans le cas d'un estimateur fondé uniquement sur un échantillon de convenance de n'importe quelle taille comportant un biais normalisé de 0,2 (par exemple, 10 points de pourcentage pour une variable dichotomique où $P = 0,5$), l'EQM sera supérieure ou égale à celle d'une estimation d'un échantillon probabiliste de 25 unités.

Remerciements

Les auteurs tiennent à remercier John Adams, Ph.D., des commentateurs qu'il nous a fournis au sujet du document manuscrit, Matthias Schonlau, Ph.D., de ses commentaires au sujet d'une version précédente du document manuscrit, Colleen Carey, B.A., Kate Sommers-Daves, B.A., et Michelle Platt, B.A. de leur collaboration à la rédaction du document.

Marc Elliott bénéficie entre autres de l'aide financière des Centers for Disease Control and Prevention (CDC U48/DP000056); les opinions exprimées dans le présent article ne reflètent pas nécessairement celles des Centers for Disease Control and Prevention.

3.5 La précision de l'estimation du biais

Dans le présent document, nous avons présumé que le biais des estimateurs de l'échantillon de convenance était connu; en pratique, le biais devra être estimé au moyen des données des deux échantillons. Nous examinons ensuite dans quelle mesure la taille de l'échantillon probabiliste limite également l'utilité de l'échantillon de convenance en raison de la nécessité d'estimer avec précision le biais résiduel.

Nous pouvons estimer e comme la différence entre la moyenne de l'échantillon probabiliste et la moyenne pondérée de l'échantillon de convenance. L'erreur type de l'estimation du biais est obtenue par $\sigma_e = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. Si $\sigma_1 = \sigma_2 = \sigma$, l'erreur type réelle de l'estimation du biais normalisée (E) est obtenue par l'équation suivante : $\sigma_E = \sqrt{1/n_1 + 1/n_2}$. Peu importe la taille de l'échantillon de convenance, ce terme ne peut jamais être inférieur à l'inverse de la racine carrée de la taille de l'échantillon probabiliste.

Il a été démontré que l'erreur relative de l'EOM pour un estimateur composite est relativement insensible aux petites erreurs des estimations du biais (Schabie 1978), ce qui est encourageant pour les biais bien estimés. Malheureusement, à moins que la TBE des échantillons probabiliste et de convenance soit élevée, l'erreur type de l'estimation de E est inutilement élevée par rapport aux valeurs de E qui rendent l'ajustement de convenance ($E < 0,10$). Par exemple, supposons qu'un échantillon probabiliste dont la TBE est de 1 000 unités et qu'un échantillon de convenance dont la TBE est de 5 000 ait une estimation un biais normalisé de 0,02 points de pourcentage. Si l'estimation ponctuelle était correcte, l'échantillon de convenance augmenterait TBE₁ de 1 667 unités. Cependant, cette estimation pourrait être de 95 % de limite de confiance supérieure), ce qui signifierait que l'augmentation est inférieure à 130.

Si nous présumons que la taille de l'échantillon de convenance sera toujours équivalente au moins au double de la taille de l'échantillon probabiliste, ces résultats signifient que pour que cette technique puisse être mise en pratique, l'échantillon probabiliste doit comporter au moins 1 000 à 10 000 unités, de manière à lever l'incertitude de l'ampleur du biais des estimateurs de l'échantillon de convenance (erreurs-types de E dans la fourchette de 0,01 à 0,04).

4. Discussion

Nous décrivons un estimateur composite, qui est une combinaison linéaire de l'estimation de la moyenne d'un échantillon probabiliste sans biais et de l'estimation de la

biaisé (pondération fondée sur les scores de propension). Nous nous servons de l'EOM de cet estimateur composite pour caractériser en fonction de la TBE les contributions de l'échantillon de convenance à un estimateur uniquement fondé sur l'échantillon probabiliste. Nous calculons ensuite la contribution maximale de l'échantillon de convenance, le rôle de la taille de l'échantillon de convenance pour approcher cette limite, et les rôles des deux tailles d'échantillon pour estimer le biais de manière suffisamment précise.

Certains spécialistes présument parfois que les petits biais des estimations des échantillons de convenance suggèrent tout autre chose. Nous démontrons que le biais normalisé des estimateurs de l'échantillon de convenance électronique, après les ajustements initiaux pour réduire le biais, doit être très petit (0,1 fois les écarts-types tout au plus, et probablement de moins de 0,05) pour que l'EOM de l'estimation globale soit significativement plus faible qu'elle ne l'aurait été sans l'échantillon de convenance. Nous démontrons en outre que des échantillons de convenance comportant des milliers ou des dizaines de milliers d'unités sont également nécessaires par souci d'efficacité. Enfin, nous démontrons la nécessité d'un gros échantillon probabiliste (de 1 000 à 10 000 unités) pour obtenir des estimations raisonnablement précises du biais résiduel des estimateurs de l'échantillon de convenance initialement corrigés en fonction du biais. Comme le biais des estimations dans une application d'enquête polyvalente a tendance à varier selon les résultats, il faut faire preuve de prudence avant de prendre la décision globale de remplacer un grand nombre d'enquêtes peu coûteuses par un plus petit nombre d'enquêtes traditionnelles.

Ce sont les grandes enquêtes qui offrent probablement la possibilité de réaliser les plus grandes économies, tout simplement en raison de leur taille. Par ailleurs, les améliorations proportionnelles de précision surviendront probablement dans les échantillons de taille intermédiaire. Les améliorations pourraient également s'avérer considérables dans les grands échantillons dont les principales inférences sont des sous-groupes plus petits. Par exemple, une enquête nationale menée auprès de 100 000 personnes pourrait faire une inférence à 200 sous-régions géographiques, comportant chacune un échantillon de 500 unités. Si l'on ajoutait à cet échantillon national un très gros échantillon de convenance électronique, estimait le biais à l'échelle nationale et admettait l'invariabilité régionale du biais, on pourrait réduire considérablement l'EOM des estimations infrarégionales en utilisant un tel estimateur composite.

Ainsi, les différences entre l'échantillon de convenance électronique et l'échantillon probabiliste non électronique ne sont pas entièrement attribuables à la sélection.

3.4 Contribution réelle de l'échantillon de convenance

Bien que la croissance maximale possible (CPMEC) soit de $1/E_2$, la croissance réelle de TEB_1 peut s'exprimer comme suit : $(k/k+1)$ CPMEC où $k = n_2E_2$. L'insuffisance de la croissance réelle de TEB_1 en raison de la CPMEC peut s'exprimer comme suit : CPMEC- $TEB_1 = 1/[(E_2)(1+n_2E_2)]$. Ainsi, les rendements de TEB_1 diminuent à mesure qu'augmente la taille de l'échantillon de convenance, la diminution se faisant plus rapide lorsque le biais est plus prononcé, étant donné que ce dernier finit par dominer toute autre réduction de la variance. La moitié de la CPMEC relevée est obtenue lorsque la TEB de l'échantillon de convenance est égale à la CPMEC. Par exemple, si le biais vaut 0,01 fois les écarts-types et qu'un échantillon de convenance a une TEB de 10 000 unités, alors la CPMEC est de 10 000, mais la contribution progressive réelle à TEB_1 équivaut à 5 000. Cette réalité suggère que les échantillons de convenance dont la TEB est de deux à 20 fois supérieure à la CPMEC seront suffisants dans la plupart des cas, ce qui représente de 67 % à 95 % de l'augmentation potentielle de la TEB. Ces connaissances heuristiques suggèrent à leur tour la collecte de 200 à 4 000 cas où E est relativement élevé ($E = 0,05$ à $0,10$), et de 5 000 à 200 000 cas où E est relativement faible ($E = 0,01$ à $0,02$). Le tableau 2 donne des exemples de TEB_1 obtenus au moyen de diverses combinaisons de tailles d'échantillons et de biais.

Tableau 2 Exemples de TEB_1 selon diverses tailles d'échantillons et niveaux de biais normalisés

n_1	n_2	E (biais normalisé %)	TEF ₁ pour TEF ₁ / m_1 &	(taille de l'éch. normalisé) ^a , l'estimation de (taille de l'éch. normalisé) ^b	probab.) de convenance)	Des estimateurs de moyennes fondés uniquement sur l'échantillon de convenance		Des estimateurs de moyennes fondés uniquement sur l'échantillon de convenance		TEF par rapport à la non-utilisation de l'échantillon de convenance.	
1 000	1 000	0,01	1 909	1 909	1 000	10 000	10 000	10 000	10 000	1,909	
1 000	1 000	0,10	1 091	6 000	1 000	10 000	10 000	10 000	10 000	1,009	
1 000	1 000	0,01	1 091	6 000	100 000	10 000	10 000	10 000	10 000	1,009	
1 000	1 000	0,10	1 091	6 000	100 000	10 000	10 000	10 000	10 000	1,010	
1 000	1 000	0,01	1 100	6 091	1 000	10 000	10 000	10 000	10 000	1,010	
1 000	1 000	0,10	1 100	6 091	1 000	10 000	10 000	10 000	10 000	1,010	
1 000	1 000	0,01	1 099	6 091	1 000	10 000	10 000	10 000	10 000	1,009	
1 000	1 000	0,10	1 099	6 091	1 000	10 000	10 000	10 000	10 000	1,009	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,01	1 000	10 000	100 000	10 000	10 000	10 000	10 000	1,000	
1 000	1 000	0,10	1 000	10 000	100 000	10 000	10				

3.2 La quantification des contributions de l'échantillon de convenance

Nous pouvons maintenant évaluer les contributions de l'échantillon de convenance en fonction du biais résiduel connu de ses estimateurs connexes. Pour ce faire, nous définirons plusieurs quantités.

$$TFE_1 = \frac{\sigma_2^2/n_1}{\varepsilon_2^2 + \sigma_2^4/n_1 + \sigma_2^2/n_2} \left(n_1 \right)$$

Supposons que

représente la taille effective de l'échantillon nécessaire pour obtenir une moyenne d'échantillon sans biais ayant la même EQM que l'estimateur composite. Pour simplifier cette expression, définissons le biais normalisé résiduel, $E = \varepsilon/\sigma_2$, et examinons le cas dans lequel les observations des populations de convenance et probabiliste ont une variance égale ($\sigma_1 = \sigma_2 = \sigma$). En pareil cas, la *croissance* de TFE_1 attribuable à l'échantillon de convenance, soit la différence entre TFE_1 avec et sans l'échantillon de convenance, est obtenue par l'équation suivante :

$$\frac{1}{1/n_2 + E^2} = n_2 \left(\frac{1 + n_2 E^2}{1} \right)$$

3.3 Contribution maximale de l'échantillon de convenance

À mesure que $n_2 \rightarrow \infty$, la croissance de TFE_1 se rapproche de $1/E^2$. Cette limite, qui représente l'inverse du biais normalisé moyen, correspond à la contribution progressive maximale possible de l'échantillon de convenance à TFE_1 (CPMEC). Si la CPMEC est faible, alors un échantillon de convenance, peu importe sa taille, ne peut pas améliorer significativement l'EQM. Si la CPMEC est significative, nous devons déterminer la taille de l'échantillon de convenance nécessaire pour obtenir une grande partie de la CPMEC.

Pour développer une intuition de l'ampleur de E (biases normalisés), nous examinons le cas important d'un résultat dichotomique, pour lequel $E = \varepsilon/\sqrt{p(1-p)}$ où p correspond à la probabilité du résultat dans la population. Le tableau 1 traduit le biais d'un résultat dichotomique, des points de pourcentage au biais normalisé, puis à la CPMEC correspondante pour $P = 0,1$ et $P = 0,5$.

Lorsque la proportion se rapproche de 50 %, un biais de 2,5 points de pourcentage limite à 400 la croissance potentielle de TFE_1 . La croissance minimale de TFE_1 pour compenser les frais associés à l'établissement du mode de déclaration électronique varie d'un utilisateur à l'autre, mais nous croyons qu'une croissance inférieure à 100 serait rarement rentable. Le tableau 1 porte à croire que les échantillons de convenance pour lesquels des biais normalisés des estimateurs limités à l'échantillon de

Nous commençons en nous posant la question suivante : Quel est l'estimateur de ce type le plus efficace lorsqu'on connaît l'ampleur du biais? Nous envisagerons ensuite la possibilité d'assouplir la supposition relative à la connaissance de l'ampleur du biais.

Supposons que n_1 et n_2 correspondent respectivement à la taille effective de l'échantillon probabiliste et à celle de l'échantillon de convenance, après division de la taille nominale des échantillons par les effets du plan de sondage associés au plan d'échantillonnage et au redressement de la non-réponse. Il faut tenir compte du score de proposition ou de tout autre pondération dans le cas de l'échantillon de convenance. La première population a une moyenne de μ , et une variance de σ_2^2 ; la deuxième a une moyenne de $\mu + e$ et une variance de σ_2^2 , où e représente le biais connu qui reste après la pondération, et μ correspond au paramètre inconnu qui nous intéresse. Les moyennes des échantillons correspondants ont une espérance μ et $\mu + e$ et une variance σ_2^2/n_1 où $i = 1, 2$ en supposant un modèle d'échantillonnage fondé sur une population infinie. Nous présumons que ces deux estimateurs ne sont pas corrélés, étant donné qu'ils proviennent d'échantillons indépendants.

D'après Rao (2003, pages 57-58), l'estimateur composite le plus efficace de μ prend la forme suivante :

$$\hat{\mu} = \frac{\bar{x}_2(\sigma_2^2/n_1) + \bar{x}_1(\varepsilon^2 + \sigma_2^2/n_2)}{\varepsilon^2 + \sigma_2^4/n_1 + \sigma_2^2/n_2}$$

avec le biais résiduel

$$\varepsilon_c = \varepsilon \left(\frac{\sigma_2^2/n_1}{\varepsilon^2 + \sigma_2^4/n_1 + \sigma_2^2/n_2} \right)$$

et

$$EQM_c = \frac{(\sigma_2^2/n_1)(\varepsilon^2 + \sigma_2^2/n_2)}{\varepsilon^2 + \sigma_2^4/n_1 + \sigma_2^2/n_2}$$

Comme on peut le voir, l'estimateur composite est une combinaison convexe des moyennes de l'échantillon de convenance et de l'échantillon probabiliste. L'influence du premier est déterminée par le ratio de l'EQM (la variance) de la moyenne de l'échantillon probabiliste par rapport à la somme de ce terme et de l'EQM de la moyenne de l'échantillon de convenance. De même, le biais résiduel correspond au biais initial multiplié par ce même ratio, tandis que EQM_c est le produit des deux EQM divisé par leur somme. Soulignons que le biais se rapproche de zéro lorsque $e \rightarrow 0$ (aucun biais de sélection dans l'estimation de l'échantillon de convenance) et lorsque $e \rightarrow \infty$ (aucun poids n'a été attribué à l'échantillon de convenance).

tables des échantillons de convenance électroniques, nous nous efforçons de repérer les situations où l'augmentation de la taille effective de l'échantillon (TEF) attribuable à l'inclusion de l'échantillon de convenance pourrait être suffisante pour justifier une approche à deux modes. Nous démontrons que les circonstances dans lesquelles un échantillon de convenance électronique supplémentaire pourrait améliorer l'EQM de façon significative sont très limitées. Bien que nous attirions sur les échantillons de convenance électroniques, la discussion qui suit s'applique également à d'autres méthodes de collecte de données peu coûteuses à faible couverture de population.

2. Problématique

2.1 Conditions initiales

Dans le cas de l'échantillon combiné probabiliste/de convenance, nous proposons de mener simultanément la même enquête auprès d'un échantillon probabiliste traditionnel (avec ou sans mode de déclaration électronique) et d'un échantillon de convenance électronique. Nous imaginons une enquête polyvalente menant à divers résultats. Dans le présent document, nous nous attarderons sur l'estimation des moyennes, mais il se pourrait que les travaux à venir appliquent ces résultats à d'autres paramètres, tels que les paramètres de régression. Au départ, nous examinerons les cas où le biais des estimations de l'échantillon de convenance est connu, mais nous évaluerons par la suite dans quelle mesure l'échantillon probabiliste permet de mesurer le biais inconnu dans chaque estimation des paramètres de l'échantillon de convenance.

Lorsque le biais est connu, on peut combiner les échantillons de convenance et probabiliste de manière à réduire l'EQM. Si les estimations de l'échantillon de convenance sont fortement biaisées, l'échantillon de convenance ne servira pas à grand-chose. Cette possibilité nécessite que l'échantillon probabiliste soit assez grand pour être considéré par lui-même. Par conséquent, une approche possible serait de réserver une petite partie du budget de l'échantillon probabiliste pour créer un gros échantillon de convenance supplémentaire.

Prenons par exemple le cas d'une enquête qui s'intéresse principalement aux estimations de l'ensemble de la population, mais pour laquelle il serait également souhaitable d'obtenir des estimations de sous-populations s'il était rentable d'obtenir un échantillon d'une taille suffisante pour assurer la précision des résultats. Supposons en outre qu'on pourrait faire 4 000 observations de probabilité et 10 000 observations de convenance pour le coût d'un échantillon probabiliste de 5 000. Pour un résultat donné, si le biais est prononcé, les erreurs-types augmentent modérément en

raison d'une faible réduction proportionnelle de la taille de l'échantillon, si le biais est généralement petit et propre à chaque sous-population, il pourrait y avoir des retombées positives sur la précision, ce qui pourrait donner lieu à des analyses de sous-populations d'une précision acceptable.

2.2 Réduction du biais initial

Nous ferons la preuve que le biais des estimateurs de l'échantillon de convenance doit être très faible pour que l'échantillon soit utile, ce qui indique qu'il pourrait être préférable de s'attarder sur les estimations de paramètres qui sont habituellement moins biaisées que les estimations des proportions ou des moyennes de la population globale non ajustées, telles que les coefficients de régression (Kish 1985).

En outre, on pourrait réduire le biais grâce au calage sur marges de l'échantillon de convenance en fonction des valeurs connues de la population (Kallon et Kasprzyk 1986), ou en appliquant les poids des scores de propension qui modélisent la sélection des membres entre les deux échantillons aux observations de l'échantillon de convenance (Rosenbaum 2002; Rosenbaum et Rubin 1983). On peut inclure une petite série d'éléments pour permettre l'utilisation d'une de ces approches. Parmi ces éléments figurent aussi bien ceux qui prédisent les différences entre les répondants qui fournissent leurs données par voie électronique et ceux qui optent pour d'autres modes de déclaration, que les éléments adaptés au contenu de l'enquête en particulier. L'effet du plan de sondage des poids des variables obtenus réduiront la TEB pour les estimateurs de l'échantillon de convenance, mais les faibles coûts de ces observations rentabilisent la compensation des effets modérés du plan de sondage. Nous pouvons ensuite estimer le biais résiduel d'un paramètre donné, comme la différence entre l'estimation de l'échantillon probabiliste et l'estimation pondérée de l'échantillon de convenance.

3. Facteurs d'efficacité

3.1 Combinaisons linéaires des estimateurs biaisés et sans biais d'une moyenne de population

Un cas spécial d'un estimateur donné dans un résultat par Rao (2003, pages 57-58) constitue l'estimateur le plus efficace des combinaisons linéaires d'échantillons de convenance et probabilistes (pondérés). Les propriétés de cet estimateur mènent à des recommandations générales au sujet des conditions de la taille de l'échantillon probabiliste, de la taille de l'échantillon de convenance et du biais de l'estimateur de l'échantillon de convenance qui permettent à l'échantillon de convenance d'améliorer significativement la TEB de l'échantillon probabiliste.

Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste

Marc N. Elliott et Amelia Haviland

Résumé

Dans le présent document, nous décrivons une méthodologie utilisée pour combiner un échantillon de convenance avec un échantillon probabiliste afin de produire un estimateur ayant une erreur quadratique moyenne (EQM) plus faible que les estimateurs fondés uniquement sur un échantillon probabiliste. Nous examinons ensuite les propriétés de l'estimateur composite obtenu, qui est en fait une combinaison linéaire des estimateurs de l'échantillon de convenance et de l'échantillon probabiliste, les poids étant fonction du biais. Nous discutons des propriétés de l'estimateur dans le contexte de l'échantillonnage de convenance électronique. Notre analyse démontre que le recours à un échantillon de convenance pour compléter un échantillon probabiliste en vue d'améliorer l'EQM de l'estimation pourrait s'avérer utile seulement dans des circonstances restreintes. Premièrement, le biais résiduel de l'estimateur fondé sur l'échantillon de convenance doit être très faible, représentant tout au plus 0,1 de l'écart-type de la population obtenu. En cas de résultat dichotomique, cela signifie un biais ne dépassant pas cinq points de pourcentage à 50 % de prévalence, et trois points de pourcentage à 10 % de prévalence. Deuxièmement, l'échantillon probabiliste devrait contenir au moins 1 000 à 10 000 observations pour donner lieu à une estimation adéquate du biais de l'estimateur de l'échantillon de convenance. Troisièmement, il doit être rentable et faisable de recueillir au moins des milliers (et probablement des dizaines de milliers) d'observations à partir de l'échantillon électronique de convenance. Les conclusions au sujet de l'utilité limitée des échantillons de convenance lorsque le biais de l'estimateur comporte un écart-type de plus de 0,1 s'appliquent également à l'utilisation directe des estimateurs en fonction de cet échantillon.

Mots clés : Biais; estimateur composite; calage.

1. Introduction

Les enquêtes en ligne (électroniques) se font de plus en plus nombreuses et diversifiées (Couper 2000). Par exemple, les échantillons probabilistes électroniques s'appuient sur une base de sondage traditionnelle et offrent la déclaration par Internet parmi d'autres modes de déclaration ou comme unique méthode. Les échantillons probabilistes électroniques peuvent avoir des taux de réponse élevés et produire des estimateurs comportant un biais de non-réponse minime (Kypri, Stephenson et Langley 2004). En revanche, les échantillons de convenance électroniques sont fondés sur les requêtes « internes » de pages Web lancées par n'importe quel internaute qui découvre le site et décide de participer (parfois en conséquence de publicité ciblant une population qui n'est pas spécifiable), ou encore, ils dépendent de la bonne volonté de panels recrutés qui ne sont pas nécessairement représentatifs de la population visée.

Le principal attrait des échantillons de convenance électroniques réside dans les coûts marginaux par cas très faibles, les coûts totaux par cas du traitement des données et la réduction des coûts de collecte de données et la réduction des coûts de publication (coûteuses, sans compter qu'elles favorisent la rapidité de la collecte de données et la réduction des coûts marginaux par cas du traitement des données). Même si l'on tient compte de quelques frais fixes, les coûts totaux par cas

Le principal attrait des échantillons de convenance électroniques réside dans les coûts marginaux par cas très faibles, les coûts totaux par cas du traitement des données et la réduction des coûts de collecte de données et la réduction des coûts de publication (coûteuses, sans compter qu'elles favorisent la rapidité de la collecte de données et la réduction des coûts marginaux par cas du traitement des données). Même si l'on tient compte de quelques frais fixes, les coûts totaux par cas

Dans le présent document, nous décrivons une méthodologie qui consiste à combiner un échantillon de convenance avec un échantillon probabiliste, en vue de produire un estimateur comportant une erreur quadratique moyenne (EQM) plus faible que celle des estimateurs fondés uniquement sur l'échantillon probabiliste. Ensuite, nous examinons les propriétés de l'estimateur composite obtenu, une combinaison linéaire des échantillons de convenance et probabiliste, les poids étant déterminés par le biais. Cette démarche nous amène à formuler des recommandations au sujet de l'utilité d'ajouter des échantillons de convenance électroniques aux échantillons probabilistes. Compte tenu des coûts différentiels très

Fienberg, S.E. (1992). Bibliographie sur la modélisation à l'aide de la saisie-réessais avec application au redressement des chiffres du recensement pour éliminer le sous-dénombrement. *Techniques d'enquête*, 18, 1, 157-169.

Hogan, H. (1993). The post enumeration survey: Operation and results. *Journal of the American Statistical Association*, 88(423), 1047-1060.

Hogan, H. (2003). L'évaluation de l'exactitude et de la couverture : théorie et conception. *Techniques d'enquête*, 29, 2, 145-156.

McLennan, W. (1997). Census of Population and Housing. Data Quality - Undercount. Australia 1996. Article d'information, 2940.0. Australian Bureau of Statistics.

Renaud, A. (2001). Methodology of the Swiss Census 2000 Coverage Survey. *Proceedings of the Survey Research Methods Section* [CD-ROM]. American Statistical Association.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la surcouverture (E-sample). Rapport de méthodes, 338-0019, Office fédéral de la statistique.

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. Rapport de méthodes, 338-0027, Office fédéral de la statistique.

Renaud, A., et Eichengraber, P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Rapport de méthodes, 338-0009, Office fédéral de la statistique.

Renaud, A., et Poterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Échantillon pour l'estimation de la sous-couverture (F-sample) et qualité du cadre de sondage des bâtiments. Rapport de méthodes, 338-0023, Office fédéral de la statistique.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Statistique Canada (1999). Couverture. Rapport technique du recensement de 1996, 92-370-XIF, Statistique Canada.

Statistique Canada (2004). Couverture. Rapport technique du recensement de 2001, 92-394-XIF, Statistique Canada.

Trewin, D. (2003). Census of Population and Housing. Data Quality - Undercount. Australia 2001. Article d'information, 2940.0. Australian Bureau of Statistics.

Wolter, K.M. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, 81(394), 338-346.

Un élément important à revoir pour de futures estimations est le choix de la population d'intérêt. Le choix de domicile économique a provoqué quelques difficultés dans les estimations car une délimitation précise de cette population était difficile. Une future estimation pourrait exclure les ménages collectifs pour éviter les problèmes pratiques de relevé mais conserver tous les types de domiciles. L'ensemble des enregistrements au domicile économique serait alors traité comme un domaine.

L'estimation des défauts de couverture d'un recensement est un projet ambitieux qui a montré son intérêt. Les résultats donnent des informations sur la qualité des données du recensement 2000 et les différents problèmes de couverture. Les prochains recensements seront essentiellement basés sur des registres. Les estimations de couverture se baseront sur l'expérience acquise lors des estimations de 2000 avec de probables adaptations pour tenir compte du nouveau système de relevé.

Remerciements

Je tiens à remercier Philippe Eichengraber du Service de méthodes statistiques de l'Office fédéral de la statistique pour les discussions fructueuses durant tout le projet. Un grand merci également à Dr. Rajendra Singh et ses collègues du Decennial Statistical Studies Division du U.S. Census Bureau pour leur assistance lors du développement des méthodes et des estimations. Merci également à toutes les personnes du recensement qui ont effectués des travaux et fourni des informations pour le bon déroulement du projet, et à Paul-André Salamin du Service de méthodes statistiques pour la relecture du papier.

Bibliographie

ABS (1997). The 1996 census of population and housing. Rapport annuel 1996-97, Australian Bureau of Statistics.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J., et Teague, A.D. (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society, Série A*, 162(2), 247-267.

6.3 Erreurs potentielles de localisation et décalage

temporel

Parmi les 49 107 éléments appartenant à l'enquête de recensement, 97,7 % sont trouvés dans l'aire de base autour de l'adresse de référence relevée durant l'enquête. Cette valeur vaut 98,1 % pour les personnes n'ayant indiqué aucun déménagement entre le jour du recensement et le jour de l'enquête. Elle vaut 83,9 % pour ceux qui ont indiqué un déménagement (1 512 personnes); voir nombres absolus dans le Tableau 5.

Il est intéressant de noter que 9,4 % des personnes du NORD n'ayant pas déménagé sont retrouvées proche de leur adresse de référence mais pas exactement dans le même bâtiment. Ces problèmes de localisation fine ont une influence négligeable sur les données du recensement. Ils montrent cependant la difficulté à identifier les bâtiments échantillonnés lors de la construction des listes des ménages sur le terrain durant l'enquête, tout comme la difficulté de l'assignation des personnes aux bâtiments durant le traitement des données du recensement. Un relevé complet-faire serait cependant nécessaire pour évaluer l'effet respectif des deux difficultés.

La localisation des personnes qui ont déménagé indique que 151 = 145 + 6 personnes sont trouvées autour de leur adresse au jour de l'enquête et non pas autour de leur ensemble de 688 personnes du NORD, parmi les 922 se trouvant dans les deux aires de base, sont en fait trouvées dans le bâtiment au jour de l'enquête. Un soin spécial ayant été mis durant l'enquête de couverture sur les questions concernant les adresses au jour du recensement et au jour de l'enquête, on considère ici que les adresses des personnes ayant déménagé sont de meilleure qualité dans les données

Tableau 5
Comparaison de la localisation des personnes appartenant à l'enquête de base ou hors de l'aire étendue pour les personnes n'ayant pas déménagé (fixes) et les personnes ayant déménagé (déménagements) entre le recensement et l'enquête

Jour enquête		déménagements			
fixes	base	base	étendue	hors étendue	total
46 689	922	69	277	1 268	
258	42	4	3	49	
648	145	6	28	179	
0	15	1	0	16	
47 595	1 124	80	308	1 512	

7. Conclusion

Les défauts de couverture globaux du recensement de la population de l'an 2000 en Suisse sont dans l'ordre de grandeur des recensements dans d'autres pays. Des spécificités apparaissent cependant au niveau des sous-groupes (par exemple régions). Parmi les trois composantes, celle de la sous-couverture est fort intéressante car elle détecte non seulement des groupes de personnes plus ou moins bien recensés mais permet également d'analyser les erreurs de localisation et de mesures. Les estimations de sur-couverture sont de leur côté limitées par le manque d'informations complémentaires au recensement pour 5^e. Elles pourraient être améliorées dans le futur par la collecte d'informations complémentaires sur les caractéristiques au jour du recensement lors d'une enquête auprès des personnes de cet échantillon (par exemple localisation et type de ménage). Les estimations de la sous-couverture nette sont basées sur plusieurs hypothèses. Les résultats dans de grands domaines semblent fiables mais certains risques, liés notamment au choix des cellules d'estimations, existent lorsque les domaines sont plus petits. Pour de futures estimations, on propose d'évaluer l'approche modèle tel qu'appliqué au Royaume-Uni au lieu des cellules d'estimations utilisées traditionnellement aux États-Unis.

dans l'enquête.

on déduit qu'au moins 151 + 688 = 839 personnes ayant déménagé sur 1 512 sont énumérées dans le recensement à une adresse qu'ils n'avaient pas au jour officiel du relevé mais quelques temps après ce jour. Le décalage exact n'est pas connu car la date du déménagement n'a pas été relevée

La mesure de l'asymétrie $\phi_x(d, d')$ prend la valeur 1,3 pour le sexe (d = homme et d' = femme). Il y a plus de femmes dans le recensement que de femmes selon l'enquête d'hommes codés hommes selon l'enquête qui sont codées femmes dans le recensement que de femmes selon l'enquête qui sont des hommes selon le recensement. La proportion d'hommes est légèrement supérieure dans l'enquête. Ces résultats doivent cependant être relativisés car ils sont basés sur très peu de cas; voir le Tableau 3. Un test de McNemar est juste significatif au seuil de 5 % sans tenir compte du plan, mais il ne le reste pas si le plan est pris en compte. On observe des asymétries par contre très nettes dans l'état civil. Il y a moins de célibataires dans l'enquête qui sont mariés dans le recensement que l'inverse (facteur 0,33 pour d = célibataire et d' = marié). De même, il y a moins de mariés dans l'enquête qui sont veuf/veuve ou divorcé/e dans le recensement que l'inverse (facteur 0,42 pour d = marié et d' = autres). L'asymétrie est également visible sur la variable du permis d'établissement. La tendance est d'avoir plus de Suisses dans l'enquête qui sont notés étrangers dans le recensement que l'inverse; de même plus de permis permanents dans l'enquête et permis temporaires dans le recensement que l'inverse (facteurs 5,22 pour d = suisse et recensement).

d' = étranger avec permis permanent et 3,83 pour d = étranger avec permis permanent et d' = étranger avec permis temporaire). Les facteurs calculés se basent sur peu de cas. Ils donnent cependant un aperçu des différences potentielles entre un relevé par le biais du questionnaire du recensement et une enquête réalisée principalement par téléphone. La variable de la vie active comporte plus de cas divergents; voir le Tableau 4. Nous avons donc par exemple moins de personnes occupées dans l'enquête de 0,46 pour d = actif et d' = non actif). Il y a également moins de personnes sans emploi dans l'enquête et non actives dans le recensement que l'inverse (facteur 0,26 pour d = sans emploi et d' = non actif). La variable de la position dans le ménage montre également des asymétries, mais ces dernières sont basées sur peu d'éléments car la dispersion des éléments dans les cases (d, d') est importante. Les variables du recensement au niveau ménage (position dans le ménage et taille de ménage) sont influencées par le processus complexe de la formation des ménages. Elles sont moins fiables que celles qui concernent les personnes. Les valeurs au niveau ménage sont plus fiables dans l'enquête.

Tableau 3 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable sexe

Enquête		Recensement	
homme	femme	total	total
24 171	24 936	49 107	776
23 967	166	24 133	204
23 967	166	24 133	204
6	0	13	13
0	6	24 564	25 319
13	6	24 564	49 883

Tableau 4 Comparaison des valeurs relevées dans l'enquête et le recensement pour la variable vie active

vie active		Enquête		Recensement	
non apparié	total	occupé	sans emploi	non actif	≤ 15 ans
total	25 163	498	14 501	8 945	49 107
apparié	23 953	188	2 007	13	26 161
occupé	300	221	323	18	845
non actif	901	89	12 143	1	13 151
≤ 15 ans	9	0	28	8 913	8 950
apparié	23 953	188	2 007	13	26 161
occupé	300	221	323	18	845
non actif	901	89	12 143	1	13 151
≤ 15 ans	9	0	28	8 913	8 950
apparié (variable imputée)	564	22	312	6	904
sans emploi	14	8	26	1	49
non actif	92	15	881	5	993
≤ 15 ans	0	0	0	0	0
total	25 587	521	14 718	9 057	49 883

faible que celui observé pour le sexe, l'âge, le permis et l'état civil. Il s'agit de la variable sur la vie active (actif, sans emploi, non actif), la position dans le ménage (seul/seule, époux/épouse, union libre, une personne avec enfant(s), autre chef de ménage, apparenté au chef de ménage, autres; résultats restreints aux ménages privés) et la taille du ménage de la personne (selon domicile économique et dans les ménages privés). Le taux R_X vaut 90,4 % pour la vie active (91,1 % parmi les valeurs non imputées), 91,4 % pour la position dans le ménage (94,9 % parmi les valeurs non imputées) et 88,3 % pour la taille du ménage.

Tableau 2 Nombre recensé C et taux estimés de sur-couverture R_{sur} , sous-couverture R_{sous} et sous-couverture nette $R_{sousnet}$ pour différents domaines [%], avec les écart-types estimés ($E-T$) correspondants

Variable	Catégories	R_{sur}	$E-T$	R_{sous}	$E-T$	$R_{sousnet}$	$E-T$
global	homme	3 497 940	0,37	1,74	0,13	1,46	0,13
	femme	3 623 686	0,33	1,55	0,10	1,37	0,13
classe âge	≤ 9	810 373	0,26	1,46	0,21	1,34	0,26
	10-19	833 185	0,27	1,30	0,19	1,04	0,22
	20-31	1 115 804	0,93	3,50	0,34	2,84	0,36
	32-44	1 544 721	0,33	1,65	0,16	1,43	0,19
	45-59	1 431 771	0,22	1,18	0,14	1,04	0,14
	60-79	1 146 709	0,10	0,91	0,13	0,82	0,12
	≥ 80	239 063	0,11	1,20	0,31	1,03	0,27
permis d'établissement	Suisse	5 674 266	0,33	1,28	0,09	0,98	0,10
	étranger permanent	1 020 242	0,33	1,85	0,29	2,89	0,32
	étranger temporaire	427 118	0,56	8,03	0,53	3,48	0,39
état civil	célibataire	2 975 643	0,50	2,07	0,18	1,72	0,19
	marité	3 377 223	0,23	1,27	0,11	1,25	0,12
	veuf/veuve	369 339	0,25	1,23	0,26	0,79	0,13
	divorcé/e	399 421	0,24	1,95	0,35	1,02	0,10
langue commune	allemand + romanche	5 128 353	0,33	1,50	0,11	1,28	0,12
	français	1 680 062	0,35	1,89	0,25	1,79	0,27
	italien	313 211	0,53	2,35	0,49	1,56	0,19
région NUTS	région lémanique	1 296 464	0,37	2,19	0,38	1,84	0,28
	espace Mittelland	1 640 489	0,35	1,39	0,15	1,25	0,10
	Nordwestschweiz	976 699	0,18	1,50	0,27	1,32	0,12
	Zürich	1 221 014	0,31	1,58	0,19	1,46	0,13
	Ostschweiz	1 020 897	0,40	1,29	0,23	1,24	0,12
	Zentralschweiz	665 904	0,36	1,57	0,25	1,19	0,12
taille commune	Ticino	300 159	0,54	2,38	0,52	1,57	0,19
	petite	1 372 958	0,34	1,50	0,15	1,12	0,14
	moyenne	2 398 256	0,41	1,32	0,16	1,07	0,19
	grande	3 350 412	0,31	2,01	0,19	1,77	0,19
type	ville	2 078 780	0,35	1,96	0,17	1,82	0,20
	agglomération	3 145 541	0,36	1,49	0,19	1,34	0,12
	rural	1 897 305	0,32	1,56	0,17	1,07	0,12
relevé	CLASSIC	265 607	0,39	1,91	0,28	1,07	0,12
	SEMI-CLASSIC	174 501	0,37	1,07	0,24	1,16	0,13
	TRANSIT + FUTURE	6 381 359	0,33	1,62	0,11	1,42	0,12
	TICINO	300 159	0,54	2,38	0,52	1,57	0,19

de détecter d'éventuels problèmes de décalage temporel entre le jour du recensement et le jour du relevé effectif des données du recensement.

6. Résultats

6.1 Estimations des défauts de couverture

Le taux global de sous-couverture nette est estimé à 1,41 % avec un écart-type de 0,12 %. Le taux de sur-couverture est de 0,35 % (écart-type = 0,03 %) et le taux de sous-couverture est de 1,64 % (écart-type = 0,11 %). Ces résultats sont dans l'ordre de grandeur de ceux d'autres pays, bien que plutôt dans les valeurs les plus basses; voir

Tableau 1.

La sur-couverture est peu importante dans la grande majorité des domaines étudiés. Le plus haut taux est observé pour les personnes entre 20 et 31 ans (0,93 % avec un écart-type de 0,09 %); voir Tableau 2. La sous-couverture est par contre élevée dans plusieurs domaines. On note par exemple un taux de 8,03 % (écart-type = 0,85 %) pour les étrangers avec des permis d'établissement temporaires (« autres permis ») et un taux de 3,50 % (écart-type = 0,50 %) pour les 20-31 ans. On note aussi un taux de sous-couverture de 2,4 % dans la région italophone du pays (langue de la commune : italien, région NUTS : Ticino, et relève : TICINO). Les résultats sont cependant liés à une relation grande variabilité (écart-type de env. 0,5 %) car les échantillons S_p et S_E ne comportent respectivement que 1 500 et 1 700 personnes dans cette région.

Tableau 1
Comparaison internationale des résultats globaux. Taux estimés de sur-couverture R_{sur} , sous-couverture R_{sous} et sous-couverture nette R_{soynet} avec écarts-type estimés correspondants. Références : Statistique Canada (1999, 2004), Hogan (1993, 2003), McLennan (1997) et Trewhin (2003)

Sur-couverture		Sous-couverture		Sous-couverture nette [%]	
	[%]		[%]		[%]
Suisse	2000	0,3	(0,03)	1,64	(0,11)
Canada	1996	0,74	(0,04)	3,18	(0,09)
Etats-Unis	1990	3,1		4,7	
	2000	-		-	
Australie	1996	0,2		1,8	
	2001	0,9		2,7	
				1,8	(0,10)

La sous-couverture nette est positive dans tous les domaines étudiés. Il n'y a donc pas de sur-couverture nette. Les plus grandes valeurs sont observées pour les étrangers avec permis permanent ou temporaire (2,89 % et 3,48 %, écart-type = 0,32 % et 0,39 %) ainsi que chez les 20-31 ans (2,84 %, écart-type = 0,36 %). Aucune différence significative n'est observée entre hommes et femmes, entre

langues et entre régions NUTS. La faible taille de l'échantillon avec la variante de relevé TICINO ne permet pas de différencier cette méthode des autres utilisées dans le pays. Des différences sont par contre significatives entre états civils, ainsi qu'entre les types et tailles de communes.

On note que le taux de sous-couverture nette est supérieur au taux de sous-couverture dans le cas des permis d'établissement permanent. Cet effet, irrégulier, est dû au choix des cellules d'estimations et au lissage qui s'en suit. La construction des cellules a en effet nécessité un regroupement des étrangers avec permis permanents et temporaires en une seule catégorie lors des agrégats permettant d'atteindre la taille minimale de 100 personnes par cellule. Par ce regroupement nous considérons les étrangers comme un groupe homogène alors qu'il ne l'est pas. Ceci montre les limites de la méthode et la difficulté de satisfaire aux hypothèses des modèles utilisés lors de l'application. Dans le cas des étrangers, on note cependant que les intervalles de confiance des taux de sous-couverture nette et de sous-couverture se recoupent. Les conséquences des faiblesses de l'application sont donc restreintes.

Notons encore que les résultats sont présentés dans des domaines définis par des variables pour lesquelles on a observé de faibles erreurs potentielles de mesure. Des résultats pour des groupes tels que définis par les caractéristiques de ménage ou de vie active ne seraient en effet que peu fiables; voir section 6.2.

6.2 Erreurs potentielles de mesure et de classification

Parmi les 49 107 éléments apparités entre l'enquête de couverture et le recensement, 96 % ne présentent aucune différence dans les 7 classes d'âge, les 3 classes d'état civil et les 3 classes de permis d'établissement (suisse, permanent, temporaire). Le taux d'appariement dans le bon domaine R_x vaut 99,3 % pour le sexe (avec et sans imputations), 98,3 % pour l'état civil (98,4 % parmi les valeurs non imputées) et 98,7 % pour le permis d'établissement (98,8 % parmi les valeurs non imputées). Le taux R_x vaut 99,5 % pour les classes d'âge (avec ou sans imputations). Il faut cependant noter que la date de naissance était, avec le nom et le prénom, une des principales variables dans l'appariement. Les différences d'âge sont donc possibles uniquement lors d'un appariement non automatique (assisté par ordinateur ou manuel). Trois variables montrent un taux d'appariement dans le bon domaine nettement plus

5. Comparaison des appartements

5.1 Erreurs potentielles de mesure

Les erreurs de mesures ou erreurs de classification sont liées aux erreurs de couverture. En effet, une personne classifiée dans le domaine d selon le recensement (par exemple personne entre 10 et 19 ans) alors qu'elle est en réalité hors du domaine (par exemple personne de 60 ans) aboutirait à une sur-couverture dans le domaine d et une sous-couverture hors de ce domaine. Cette erreur de classification (« misclassification ») ne provoque pas d'erreur de couverture au niveau global mais une erreur au niveau de sous-groupes de la population.

Les raisons des différences entre les valeurs relevées dans deux enquêtes telles que le recensement et l'enquête de couverture peuvent être très diverses et complexes à dissocier. Il faut en effet compter sur des erreurs d'apartemen- ment, des différences résultant des méthodes de relève (questionnaire papier ou téléphone/face-à-face) et du traitement des données, ou encore sur des différences réelles dues au décalage temporel entre les relevés (décembre 2000 ou avril-mai 2001). De plus, il est difficile de déterminer la réponse correcte si on a deux valeurs différentes : le recensement ? l'enquête ? une autre valeur pas relevée ?

Les erreurs de mesures potentielles des données du recensement sont analysées sur la base de l'ensemble des appartements entre l'échantillon indépendant S_p et le recensement. On choisit de déterminer quelles sont les variables qui montrent respectivement peu ou beaucoup de problèmes potentiels de classification, sans faire un jugement sur la qualité de l'un ou l'autre des relevés. Ces informations sont notamment utiles pour évaluer le choix des cellules d'estimations pour le système dual et choisir les sous-groupes pour lesquels les estimations des défauts de couverture sont les plus fondées.

On définit, pour la variable catégorielle X , le taux d'appartenance dans le bon domaine R_X comme suit :

$$R_X = \frac{\sum_{j \in S_p} w_{p,j} R_{X,j}}{\sum_{j \in S_p} w_{p,j}} \quad (11)$$

avec $w_{p,j}$ le poids de l'élément j de l'échantillon S_p appartenant (S_p match) et le statut de classification $R_{X,j}$ qui vaut 1 si l'élément j se trouve dans la même classe dans le recensement et dans l'enquête, et 0 sinon. La valeur de R_X est estimée avec l'ensemble des éléments appartenés et avec le sous-groupe des éléments sans imputation dans le recensement.

On définit également une mesure d'asymétrie $\phi_X(d, d')$ pour les classes d et d' de la variable X :

$$\phi_X(d, d') = \frac{\sum_{j \in S_p} w_{p,j} I_j^f(d, d')}{\sum_{j \in S_p} w_{p,j} I_j^f(d', d)} \quad (12)$$

avec $I_j^f(d, d') = 1$ si l'élément j se trouve dans le domaine d selon l'enquête et dans le domaine d' selon le recensement, et 0 sinon. Le facteur $\phi_X(d, d')$ vaut 1 s'il y a équilibre dans les erreurs de classification, c'est-à-dire si le nombre d'éléments dans d selon l'enquête et dans d' selon le recensement est égal au nombre dans d' selon l'enquête et dans d selon le recensement. Plus il s'éloigne de 1, plus l'équilibre est rompu.

5.2 Erreurs potentielles de localisation

Les comparaisons entre le recensement et l'enquête permettent également d'étudier la localisation géographique des personnes. Dans les données du recensement, nous avons une unique adresse si la personne a un unique domicile et deux adresses - principale et secondaire - si la personne a deux domiciles. Dans les données de l'enquête, nous avons une ou deux adresses au jour du recensement, une ou deux adresses au jour de l'enquête et l'information sur un éventuel déménagement entre les deux dates. Si une personne a un unique domicile et n'a pas déménagé, son adresse principale au jour du recensement et son adresse principale au jour de l'enquête sont identiques. Elle n'a pas d'adresses secondaires.

Différentes mesures de distance sont envisagées pour déterminer les erreurs potentielles de localisation dans le recensement. Pour des raisons pratiques, notamment de données à disposition, nous définissons des adresses géographiques autour de l'adresse principale de la personne telle que relevée durant l'enquête pour le jour du recensement (*adresse de référence*). Elles sont définies à partir des numéros postaux relevés durant l'enquête. L'aire de base de la personne est définie par l'ensemble des communes qui possèdent des bâtiments dans le numéro postal de son adresse de référence. La définition de cette aire utilise les données du registre suisse des bâtiments car ce dernier possède une information sur l'adresse postale et la commune des bâtiments. L'aire étendue comporte les communes de l'aire de base et l'ensemble des communes qui leurs sont adjacentes; voir Renaud (2004) pour des exemples.

De façon similaire aux erreurs de classification, les erreurs de localisation ne provoquent pas d'erreurs de couverture au niveau global mais des erreurs au niveau de sous-groupes tels des régions ou des types de communes. Différents taux peuvent être définis. On retiendra le taux de localisation de base et le taux de localisation étendue; tous deux pondérés par $w_{p,j}$ le poids de l'élément j de l'échantillon S_p appartenant. Le statut de localisation prend la valeur 1 si l'élément est trouvé dans l'aire de base ou respectivement étendue, et 0 sinon. On étudiera en particulier la localisation des personnes qui ont déménagé, afin

Cette forme de correction est préférée au quotient entre la somme des poids sans l'unité primitive α car elle permet de prendre en compte la variabilité due au nombre inconnu d'éléments dans la strate.

L'estimateur du jackknife devient :

$$\theta^{JK} = \sum_{m_2} \sum_{\alpha=1}^h \theta_{m_2}^{(\alpha)} - \sum_{m_1} \theta_{m_1}^{(h)}$$

avec les pseudo valeurs $\theta_{m_2}^{(\alpha)} = m_2 - (m_2 - 1)\theta_{m_2}^{(h)}$. L'estimateur de sa variance peut prendre différentes formes; voir par exemple Shao and Tu (1995). Nous appliquons la forme suivante :

$$v(\theta^{JK}) = \sum_h \frac{m_h}{m} - \frac{1}{\sum_{m_2} \theta_{m_2}^{(\alpha)} - \sum_{m_1} \theta_{m_1}^{(h)}} - \theta^{(h)2}, \quad (10)$$

avec $\theta^{(h)} = \sum_{m_2=1}^{\alpha=1} \theta_{m_2}^{(\alpha)} / m^{(h)}$. Finalement, on utilise $v(\theta^{JK})$ comme estimateur de la variance de $\hat{\theta}$. Les estimations dans des sous-groupes utilisent la même forme de l'estimateur avec l'intégration d'un indicateur de domaine dans la construction des $\theta_{m_2}^{(\alpha)}$. Aucune correction pour la population finie n'est appliquée dans les estimations. De plus on ne tient pas compte d'autres variabilités telle que celle induite par le modèle de pondération pour la non-réponse dans S^{sp} .

Des problèmes, tels le manque de stabilité de l'estimation dans les strates avec peu d'unités primaires, sont apparus durant l'application. Les essais de partage de certaines unités primaires et une comparaison avec la linéarisation de Taylor ou un jackknife simple permettent cependant de penser que les estimateurs de variance par le jackknife stratifé présentés dans ce document sont plutôt conservateurs.

4. Choix des statuts d'appariement et d'énumération corrects

Un élément clé des estimations de couverture est la définition de statut d'appariement correct pour les éléments de S^p et du statut d'énumération correct pour les éléments de S^E . Ces statuts corrects sont définis à partir des statuts de bases E_j et F_j déterminés durant les appariements.

Un appariement avec un élément du recensement faisant partie d'un ménage collectif est-il accepté comme appariement correct pour un élément de S^p ou s'agit-il d'une sous-couverture de la population d'intérêt? Un double hors de la population d'intérêt pour un élément de S^E est-il vraiment considéré comme un double, et donc une sur-couverture, ou devrait-il être exclu? Une définition claire s'impose. De plus, les statuts utilisés dans les estimations de la sous-couverture nette doivent être choisis de manière à satisfaire l'équilibre entre la sur- et la

pas de données de référence complémentaires au recensement. Pour les estimations de sous-couverture nette, il est important de satisfaire la contrainte d'équilibre. Les critères utilisés dans la définition des statuts corrects sont donc la complétude, la pertinence et l'unicité. Les critères de l'appartenance à la population d'intérêt et de la localisation ne peuvent pas être considérés car ils ne sont pas utilisables dans la définition du statut d'énumération correct. Les critères de complétude, pertinence et unicité sont déjà intégrés dans la construction des statuts de base E_j et F_j . On fait donc les estimations avec les statuts de base E_j et F_j . Pour les estimations n'utilisant pas le système dual et le besoin d'équilibrage, il est possible d'utiliser d'autres critères pour définir les statuts corrects. D'autres types de statuts d'appariements corrects sont notamment utilisés dans l'analyse des erreurs potentielles de mesure de la section 5 énumérations présentées dans Renaud (2004).

Les critères de définition des statuts corrects sont de sur-couverture).

recensement qui ont été appariés à des éléments de S^p permettent d'identifier les personnes (complétude) et que ces personnes devaient bien être recensées (pertinence). On estime aussi qu'ils sont uniques car l'unicité, bien que pas contrôlée pour les appariements, est réalisée dans la grande majorité des cas contrôlés dans S^E . Les critères d'appartenance à la population et de localisation sont contrôlés par comparaison avec les informations relevées dans l'enquête de couverture; considérées comme référence. Aucun relevé de couverture n'était organisé pour régler les cas peu clairs. Pour ce qui concerne S^E , nous avons à disposition le critère de complétude considéré comme respecté dans les données du recensement et les résultats concernant l'unicité et la pertinence obtenus dans l'appariement avec le reste du recensement. Pour les doubles et triples, on définit $E_j = I/d'$, avec $d' =$ nombre de doubles/triples dans la population d'intérêt selon le recensement. Les critères d'appartenance à la population d'intérêt et de localisation des éléments de S^E ne peuvent être contrôlés car nous n'avons pas de données de référence complémentaires au recensement.

sement. Pour les estimations de sous-couverture nette, il est important de satisfaire la contrainte d'équilibre. Les critères utilisés dans la définition des statuts corrects sont donc la complétude, la pertinence et l'unicité. Les critères de l'appartenance à la population d'intérêt et de la localisation ne peuvent pas être considérés car ils ne sont pas utilisables dans la définition du statut d'énumération correct. Les critères de complétude, pertinence et unicité sont déjà intégrés dans la construction des statuts de base E_j et F_j . On fait donc les estimations avec les statuts de base E_j et F_j . Pour les estimations n'utilisant pas le système dual et le besoin d'équilibrage, il est possible d'utiliser d'autres critères pour définir les statuts corrects. D'autres types de statuts d'appariements corrects sont notamment utilisés dans l'analyse des erreurs potentielles de mesure de la section 5 énumérations présentées dans Renaud (2004).

Ainsi une sous-couverture nette nulle dans un domaine ne signifie pas qu'aucun défaut de couverture existe dans ce domaine.

Les estimations proposées reposent sur les hypothèses du modèle dual, le choix des cellules d'estimations, et le choix des statuts définissant les estimateurs $R_{c,k}$ et $R_{m,k}$. Le modèle dual est intéressant car il prend en compte le fait que certaines personnes ne sont atteintes ni par le recensement (capture) ni par l'enquête de couverture (recapture). Cependant, une série de contraintes doivent être respectées afin d'éviter des biais d'estimation. L'enquête de couverture et le recensement doivent être totalement indépendants. L'appariement doit être de très bonne qualité. Le modèle doit être appliqué dans des cellules avec des personnes ayant la même probabilité d'être énumérées dans le recensement, respectivement dans l'enquête; voir la section 3.3.

Enfin, la population ne doit pas trop changer entre le jour du recensement et celui de l'enquête. De leur côté, les estimateurs $R_{c,k}$ et $R_{m,k}$ sont basés sur la qualité de l'appariement et de la recherche des enregistrements erronés. De plus, il s'agit de faire en sorte que la définition d'un appariement correct dans S_p et celle d'un enregistrement correct dans S_e soient identiques, *i.e.*, équilibre entre sur- et sous-couverture (« balancé »); voir la section 4. Tous ces éléments sont pris en compte dans la mesure du possible dans les présentes estimations.

L'estimation de la sous-couverture nette dans un domaine d a la forme $R_{\text{sous},d} = 1 - R_{\text{nc},d} = 1 - C_d / N_d$ avec C_d le nombre recensé dans le domaine et N_d l'estimation du vrai total. L'estimation du vrai total N_d est basée sur un modèle nommé *synthétique* qui suppose que le facteur de correction est fixe dans chaque cellule $k = 1, \dots, K$:

$$(6) \quad N_d = \sum_k N_{k,d} = \sum_k C_{k,d} F_k.$$

$C_{k,d}$ est le nombre recensé dans la population d'intérêt dans l'intersection entre la cellule k et le domaine d et F_k est le facteur de correction de la couverture dans la cellule k . L'hypothèse du modèle synthétique est respectée si le comportement de tout sous-ensemble dans la cellule est identique à celui de la cellule entière. Cette homogénéité doit être contrôlée au mieux par le choix des cellules. On reprend ici les cellules homogènes définies pour le modèle dual.

3.3 Cellules d'estimation

Les cellules d'estimation $k = 1, \dots, K$ sont constituées de façon à grouper les éléments ayant des probabilités d'énumération homogènes dans le recensement, respectivement dans l'enquête, (hypothèse dual) et des taux de couverture nette homogènes (hypothèse synthétique). On

3.4 Variance des estimateurs de couverture

Renaud (2004) pour plus de détails.

Au final, nous obtenons 121 cellules d'estimations; voir trop petites (langue officielle de la commune en 2 catégories, classe d'âge en 7 catégories et sexe en 2 catégories). Les trois variables les plus influentes (variable binaire : P_j). Les trois variables les plus influentes sont croisées : nationalité en 2 catégories, état civil en 2 catégories et tailles de la commune en 3 catégories. Les autres variables sont ensuite intégrées successivement en faisant des regroupements lorsque les tailles de cellules sont trop petites (langue officielle de la commune en 2 catégories, classe d'âge en 7 catégories et sexe en 2 catégories). Au final, nous obtenons 121 cellules d'estimations; voir

La variance des estimateurs est estimée par un jackknife stratifié appliqué sur les unités primaires - identiques - de S_p et S_e . On note que la variance de la sous-couverture estimée $R_{\text{sous}} = 1 - R_{\text{nc}}$ est égale à la variance du taux d'appariement estimé R_m . De même la variance de la sous-couverture $R_{\text{nc}} = 1 - R_c$ est égale à celle du taux d'enregistrement correct R_c et la variance de la sous-couverture nette $R_{\text{sous-net}} = 1 - R_{\text{nc}}$ est égale à celle du taux de couverture nette $R_{\text{sous-net}}$.

Soit θ le paramètre d'intérêt prenant la forme d'une moyenne pondérée de statuts dans le cas de la sous-couverture et de la sur-couverture, et la forme d'une fonction linéaire de quotients entre deux moyennes pondérées dans le cas de la sous-couverture nette. Son estimateur est $\hat{\theta}$.

Soient $h = 1, \dots, H$ la strate utilisée au premier degré de l'échantillonnage, $i = 1, \dots, m_h$ le numéro de l'unité primaire dans la strate h (numéro postal pour NORD ou commune pour TICINO), et $j = 1, \dots, m_{hi}$ le numéro de la personne dans l'unité primaire i de h . Pour les besoins du jackknife, les échantillons S_p et S_e sont partitionnés, dans chaque strate h , en m_h sous-ensembles correspondant aux personnes dans les unités primaires $\alpha = 1, \dots, m_{hi}$.

Soit $\hat{\theta}_{(h\alpha)}$ l'estimateur ayant la même forme que $\hat{\theta}$ mais calculé sur l'échantillon auquel on a retiré l'unité primaire α de la strate h . On note que les estimateurs $R_{m(h\alpha),k}$ et $R_{c(h\alpha),k}$ sont combinés pour former $R_{\text{net}(h\alpha),k}$.

$$(7) \quad R_{\text{net}(h\alpha),k} = C \left[\sum_{k=1}^K C_k R_{c(h\alpha),k} - R_{c(h\alpha),k} \right].$$

Les poids corrigés w_{hij} utilisés pour le calcul des $R_{m(h\alpha),k}$ et $R_{c(h\alpha),k}$ ont la forme :

$$(8) \quad w_{hij} = \begin{cases} 0 & \text{si } i = \alpha \\ w_{hij} m_h^{-1} & \text{si } \alpha \in h \text{ et } i \neq \alpha \\ w_{hij} & \text{si } \alpha \notin h \end{cases}$$

3.2 Couverture nette

Le *taux de sous-couverture nette* est estimé par $R_{\text{sous-net}} = 1 - R_{\text{net}}$ avec $R_{\text{net}} = C/N$ l'estimation du *taux de couverture nette*, C le nombre recensé dans la population d'intérêt et N l'estimation du vrai total dans la population d'intérêt. Si $R_{\text{sous-net}}$ est négatif, nous sommes en présence d'une sur-couverture nette.

L'estimation du vrai total N est basée sur le modèle dual (Wolter 1986). Ce modèle repose sur un principe de capture (recensement) et recapture (enquête de couverture). Il est appliqué dans des cellules d'estimation $k = 1, \dots, K$ afin de satisfaire au mieux les hypothèses du modèle; voir discussion plus bas. Ainsi, l'estimation du vrai total N est composée de la somme des vrais totaux estimés N_k dans des cellules d'estimations disjointes recouvrant l'ensemble de la population d'intérêt $k = 1, \dots, K$:

$$N = \sum_{k=1}^K N_k \quad (3)$$

Les totaux estimés N_k ont la forme donnée par le modèle dual :

$$\hat{N}_k = [N_{1+k}] \begin{bmatrix} N_{1+k} \\ N_{1+k} \end{bmatrix}, \quad (4)$$

avec N_{1+k} le total des enregistrements correctement comptés dans la cellule k durant la capture (recensement), N_{1+k} le total dans k durant la recapture (estime sur la base de l'échantillon S_p) et N_{1+k} le nombre d'enregistrements communs aux deux listes (estime sur la base des appartements entre S_p et le recensement).

Les différents termes de l'équation (4) sont estimés sur la base des estimations de sous-couverture et de sur-couverture. Il s'agit d'une extension du modèle de Wolter (1986) proche de celle utilisée notamment par Hogan (2003). Ainsi, le total des enregistrements correctement comptés dans le recensement N_{1+k} est estimé par le produit du total recensé C_k par le taux d'enregistrements corrects R_c afin de tenir compte de la sur-couverture. De plus, le rapport entre le total dans la recapture N_{1+k} et le nombre d'enregistrements communs aux deux listes N_{1+k} est estimé par l'inverse du taux d'appariement $R_{m,k}$ entre l'enquête de couverture et le recensement afin de tenir compte de la sous-couverture. On obtient :

$$\hat{N}_k = [C_k \hat{R}_c][\hat{R}_{m,k}^{-1}] = C_k \hat{F}_k \quad (5)$$

avec $\hat{F}_k = \hat{R}_c \hat{R}_{m,k}^{-1}$ le *facteur de correction de la couverture* dans la cellule k . Le facteur \hat{F}_k combine les effets de sur-couverture et de sous-couverture de la cellule k estimés sur la base des échantillons S_p et S_r . On note qu'une sous-couverture dans un domaine peut être compensée par une sur-couverture dans le même domaine.

est partie à l'étranger). Un élément j est considéré unique si aucun double/triple n'est détecté dans le recensement. Aucune interview complémentaire n'a lieu auprès du S_r . Il n'y a donc pas d'information complémentaire au recensement sur les personnes de S_r (localisation réelle ? type d'habitat ?). La recherche des doubles/triples et des cas suspects aboutit à un *statut d'énumération* E_j pour chaque élément j de S_r . Le statut E_j vaut 1 si l'élément devait bien être énuméré dans le recensement (valeur par défaut) et 0 s'il ne devait pas l'être. En pratique, il peut prendre des valeurs entre 0 et 1 si le cas n'est pas déterminé de manière précise. Ainsi, les doubles et triples reçoivent respectivement les valeurs 1/2 et 1/3 s'il n'y a pas d'informations permettant de déterminer l'enregistrement correct parmi les enregistrements détectés. Ces cas, rares, correspondent à des personnes ayant rempli plusieurs questionnaires au recensement sans qu'un lien n'ait été fait entre eux durant le traitement des données.

3. Estimateurs de la couverture

3.1 Sous-couverture et sur-couverture

Le *taux de sous-couverture* est estimé par $R_{\text{sous}} = 1 - R_m$ où R_m est l'estimation du *taux d'appariements corrects* basée sur l'échantillon S_p . De façon similaire, on définit le *taux de sur-couverture* $R_{\text{sur}} = 1 - R_c$ avec R_c l'estimation du *taux d'enregistrements corrects* basée sur l'échantillon S_r . Les taux d'appariements corrects et d'enregistrements corrects sont estimés par les moyennes pondérées des statuts d'appariement P_j et d'énumération E_j :

$$\hat{R}_m = \frac{\sum_{j \in S_p} w_{P,j} P_j}{\sum_{j \in S_p} w_{P,j}} \quad \text{et} \quad \hat{R}_c = \frac{\sum_{j \in S_r} w_{E,j} E_j}{\sum_{j \in S_r} w_{E,j}} \quad (1)$$

avec $w_{P,j}$ le poids de l'élément j de l'échantillon S_p et $w_{E,j}$ le poids de l'élément j de l'échantillon S_r . Nous notons que le dénominateur de \hat{R}_c est la somme des poids $w_{E,j}$ de S_r et non pas le nombre C d'enregistrements connus dans le recensement afin d'avoir un estimateur potentiellement moins biaisé.

L'estimation des taux de sous-couverture et de sur-couverture dans un domaine d est donnée par $R_{\text{sous},d} = 1 - \hat{R}_{m,d}$ et $R_{\text{sur},d} = 1 - \hat{R}_{c,d}$ avec

$$\hat{R}_{m,d} = \frac{\sum_{j \in S_p} w_{P,j} P_j I_{jd}}{\sum_{j \in S_p} w_{P,j} P_j I_{jd}} \quad \text{et} \quad \hat{R}_{c,d} = \frac{\sum_{j \in S_r} w_{E,j} E_j I_{jd}}{\sum_{j \in S_r} w_{E,j} E_j I_{jd}} \quad (2)$$

Les identificateurs I_{jd} et J_{jd} prennent la valeur 1 si l'élément j , respectivement de S_p et S_r , se trouve dans le domaine d et la valeur 0 sinon.

l'échantillon d'environ 16 000 bâtiments avant de passer à un sous-échantillonage de bâtiments permettant d'atteindre un total de environ 27 000 ménages. Pour plus d'information sur l'échantillonage et la procédure d'enquête, voir Renaud (2001) et, plus détaillé, Renaud et Eichenberger (2002).

L'enquête de couverture consiste à contacter les 27 000 ménages; par téléphone si un numéro est trouvé et en face-à-face si cela n'est pas le cas. On relève les variables permettant un appariement avec le recensement et la définition de sous-groupes intéressants pour l'étude de la couverture (variables sociodémographiques, adresses). Le relevé porte sur tous les membres de tous les ménages des bâtiments sélectionnés.

L'échantillon final S_p contient $n_p = 49\,883$ personnes dans la population d'intérêt (domicile économique et ménage privé). 88 % des ménages ont été atteints par téléphone et 12 % en face-à-face. La pondération dépend de l'échantillonage et d'un ajustement pour la non-réponse. L'ajustement pour la non-réponse repose sur un modèle d'homogénéité dans des cellules construites sur la base des services d'échantillonage et de la connaissance ou non de l'existence d'un numéro de téléphone (intervenus prévu par téléphone ou en face-à-face). Il intègre également une estimation de la proportion de vrais ménages parmi les ménages à contacter. Une partie non négligeable des ménages à contacter étaient en effet formée de logements de vacances, de commerces ou encore d'entreprises. Aucun calage n'est appliqué car les données auxiliaires disponibles ne sont pas indépendantes du recensement. Il n'y a pas de documents dans Renaud et Potier (2004).

Sur la base des questions posées durant l'enquête et des divers contrôles de plausibilité, on fait l'hypothèse que les données de S_p sont correctes et utilisables pour l'appariement avec le recensement. Les critères de qualité utilisés sont les suivants :

- *complétude* : l'enregistrement permet d'identifier la personne;
- *pertinence* : la personne devait bien être recensée;
- *unicité* : la personne est listée une seule fois;
- *appartenance à la population d'intérêt*, la personne est listée au domicile économique et dans un ménage privé;
- *localisation* : la personne est listée à la bonne adresse au jour du recensement.

L'appariement entre l'échantillon S_p et le recensement permet de déterminer le statut d'appartenance P_j de chaque élément j de S_p . Le statut P_j vaut 1 si l'élément est apparié dans le recensement (personne recensée) et 0 si cela n'est pas le cas (personne non recensée). Dans notre cas, les

2.3 Échantillon S_E et recherche des enregistrements erronés (sur-couverture)

L'objectif de taille de l'échantillon S_E est fixé à environ 55 000 personnes. Cette valeur, un peu supérieure à celle de S_p , n'influe que peu sur le traitement des données car il n'y a pas de travail sur le terrain ni d'interview complémentaire au recensement.

L'échantillon S_E est sélectionné dans les données du recensement selon un tirage à deux degrés. Seuls les éléments faisant partie de la population d'intérêt sont éligibles (enregistrements au domicile économique sans les membres des ménages collectifs). Les unités primaires de S_E sont identiques aux unités primaires de S_p (numéros postaux et communes). La liste des numéros postaux du plan NORD utilisée pour S_p ne correspond cependant pas exactement à la liste des numéros postaux présents dans les données du recensement. Les enregistrements du recensement se trouvant dans des numéros postaux inexistant dans la liste utilisée pour S_p sont donc redistribués dans des numéros existants en tenant compte de la localisation géographique (affectation de numéros postaux fictifs pour l'échantillonage). Au deuxième degré, on tire des enregistrements de la population d'intérêt selon un plan aléatoire simple, sans degrés intermédiaires. L'allocation est choisie de façon à obtenir des poids constants dans les strates d'échantillonage des unités primaires. Au final, l'échantillon comporte $n_E = 55\,375$ enregistrements (Renaud 2003).

On fait l'hypothèse que les enregistrements de S_E permettent d'identifier les personnes (complétude) car il y a peu d'imputation dans les données du recensement et que la plupart des questionnaires étaient préimprimés à partir des registres des habitants. La pertinence et l'unicité sont déterminées lors d'un appariement entre S_E et le reste du recensement selon un procédé similaire à l'appariement entre S_p et le recensement. Dans notre cas, il s'agit d'une recherche de doublets ou de triplets des éléments de S_E , complétée par une analyse de cas suspects dans S_p . Un élément j est considéré comme pertinent s'il n'est pas considéré comme erroné dans l'analyse des suspects (par exemple note sur le questionnaire indiquant que la personne

2. Les trois jeux de données

2.1 Recensement

Le recensement de l'an 2000 a été réalisé sous l'égide de l'Office fédéral de la statistique, avec la date de référence du 5 décembre 2000. Des informations ont été relevées pour 7,3 millions d'habitants, 3,1 millions de ménages, 3,8 millions de logements et 1,5 millions de bâtiments. Les différents niveaux ont ensuite été liés par des identificateurs communs lors du traitement des données.

Le relevé des personnes et ménages était sous la responsabilité des 2 896 communes politiques suisses. Ces dernières avaient le choix entre différentes formes de

relève :

- CLASSIC : agents recenseurs;
- SEMI-CLASSIC : préimpression de questionnaires sur la base du registre communal des habitants, envoi par la poste et collecte par des agents recenseurs;
- TRANSIT : préimpression de questionnaires, envoi et retour par la poste;
- FUTURE : idem à TRANSIT avec liens entre les ménages et les logements fournis par la commune;
- TICINO : similaire à TRANSIT mais restreint au canton du Tessin.

La majorité des communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO offraient également la possibilité de remplir les questionnaires sur Internet. Les 2 208 communes SEMI-CLASSIC, TRANSIT, FUTURE et TICINO qui ont utilisé la préimpression des questionnaires sur la base des registres communaux des habitants contiennent près de 96 % de la population. Les travaux d'envois et de contrôle des retours de la plupart de ces communes étaient organisés dans un centre national.

Le jeu de données des personnes comporte 7 452 075 entrées. Il a la particularité de comporter deux enregistrements pour la même personne si cette dernière a deux domiciles (2,3 % de la population, par exemple étudiant avec un domicile chez ses parents et un domicile proche de son école). Dans le cas de deux domiciles, l'un est codé comme *domicile économique* et l'autre comme *domicile civil*. Le domicile économique est l'endroit où la personne passe le plus de temps par semaine et le domicile civil est l'endroit où la personne a ses papiers officiels (acte de naissance pour les Suisses, permis de séjour pour les étrangers). Dans le cas d'un unique domicile, ce dernier est le domicile économique et le domicile civil tout à la fois. La *population résidente* en Suisse de 7 280 010 personnes est définie par l'ensemble des enregistrements au domicile économique.

Les ménages sont classifiés en *ménages privés*, *collectifs* et *administratifs*. Les ménages privés sont par exemple des

Les données du recensement ne comportent aucune imputation au niveau des enregistrements car les communes ont envoyé les informations de base pour les non-répondants (unit nonresponse). Des valeurs sont cependant imputées dans les cas de données manquantes ou d'incohérence dans les questionnaires (item nonresponse).

La *population d'intérêt* pour les estimations de couverture est la population résidente (domicile économique) dans des ménages privés et administratifs. Les ménages collectifs, qui représentent 2,3 % de la population résidente recensée, sont exclus des estimations.

2.2 Échantillon S_p , enquête de couverture et appariement (sous-couverture)

L'objectif de l'échantillon S_p est fixé à environ 50 000 personnes. Faute de bases existantes en Suisse, cette valeur a été déterminée de manière approximative sur la base des expériences faites à l'étranger. Les résultats austro-liens de 1996 ont notamment été utilisés car le plan d'échantillonnage de leur enquête de couverture était similaire à celui prévu en Suisse en 2000 (ABS 1997).

L'échantillon S_p , indépendant du recensement, est construit en deux parties : le canton du Tessin (TICINO) et le reste de la Suisse (NORD). Les deux parties utilisent un tirage à plusieurs degrés. Le premier degré consiste en la sélection de 303 unités primaires, communes politiques pour TICINO et numéros postaux pour NORD, selon un plan stratifié avec un tirage proportionnel au nombre de bâtiments. Le deuxième degré consiste en un tirage aléatoire simple d'un nombre fixe de 60 bâtiments par unité primaire. Dans le plan NORD, ces bâtiments sont répartis dans un maximum de 3 communes de distribution du courrier grâce à un degré d'échantillonnage intermédiaire. L'échantillonnage est ainsi construit de manière à regrouper le travail sur le terrain tout en limitant la variabilité des poids. Pour des questions pratiques et de ressources, les numéros postaux comportant une grande proportion de bâtiments sans adresses postales complètes ou codés comme inhabités sont sélectionnés avec une probabilité plus faible que les autres numéros postaux. Il s'agit principalement de numéros postaux dans des régions rurales ou encore de zones industrielles, peu enclins à des défauts de couverture importants. Des listes exhaustives de ménages sont établies sur le terrain avec l'aide des employés postaux dans

Estimation de la couverture du recensement de l'an 2000 en Suisse : méthodes et résultats

Anne Renaud¹

Résumé

Les défauts de couverture sont estimés et analysés pour le recensement de la population de l'an 2000 en Suisse. La composante de sous-couverture est estimée sur la base d'un échantillon indépendant du recensement et d'un appartement avec le recensement. La composante de sur-couverture est estimée sur la base d'un échantillon tiré dans la liste du recensement et d'un appartement avec le reste du recensement. Les composantes de sur- et sous-couverture sont ensuite combinées pour obtenir une estimation de la couverture nette résultante. Cette estimation est basée sur un modèle de capture-recapture, nommé système dual, combiné avec un modèle synthétique. Les estimateurs sont calculés pour la population entière et différents sous-groupes, avec une variance estimée par un jackknife stratifié. Les analyses de couverture sont complétées par une étude des appariements entre l'échantillon indépendant et le recensement afin de déterminer les erreurs de mesure et de localisation potentielles dans les données du recensement.

Mots clés : Recensement; erreurs de couverture; système dual; plan d'échantillonnage à plusieurs degrés; erreurs de mesure.

1. Introduction

Dans tout recensement, certaines personnes ne sont pas recensées alors qu'elles devraient l'être et d'autres sont comptées deux fois ou n'auraient pas dû être recensées. Il y a donc de la sous-couverture et de la sur-couverture, dont le bilan est très souvent une sous-couverture nette. La sous-couverture nette est par exemple estimée à 1,6 % aux États-Unis en 1990 (Hogan 1993), 2,2 % au Royaume-Uni en 1991 (Brown, Diamond, Chambers, Buckner et Teague 1999) et 3 % au Canada en 2001 (Statistique Canada 2004). Aux États-Unis en 2000, la couverture nette correspond par contre à une sur-couverture de 0,5 % (Hogan 2003). Les défauts de couverture peuvent fortement varier entre sous-groupes de la population. Aux États-Unis en 2000, on note par exemple que les Noirs avaient une sous-couverture nette de 1,8 % alors que les Blancs avaient une sur-couverture de 1,1 %. De plus, les valeurs varient souvent entre classes d'âge et régions par exemple. Ces défauts de couverture, et les autres erreurs telles les erreurs de mesure, conduisent à une image biaisée de la population. Elles sont donc étudiées afin d'avoir une information sur la qualité des données disponibles et trouver des pistes pour améliorer les relevés censitaires auprès de la population.

Le recensement de la population de l'an 2000 en Suisse donne une image de la population au 5 décembre 2000. Pour la première fois, on estime les défauts de couverture et la couverture nette résultante du recensement 2000 sont toutes trois analysées. La sous-couverture est estimée sur la base d'un échantillon de personnes S_p ,

indépendant du recensement, sur lequel on organise une enquête de couverture quelques mois après le recensement (relevé entre avril et mai 2001). Les données de cette enquête sont alors appariées avec les données du recensement afin de déterminer si la personne de S_p a été recensée ou pas. La sur-couverture est estimée sur la base d'un échantillon de personnes S_r tiré parmi les enregistrés du recensement. Une recherche de doubles et d'autres enregistrements erronés permet alors de déterminer si l'enregistrement correspond à une vraie personne à recenser ou pas. La couverture nette est estimée sur la base d'un modèle de capture-recapture nommé système dual (Wolter 1986, Fienberg 1992). L'estimateur dual est appliqué dans des cellules homogènes et les résultats recombinés en suivant un modèle synthétique pour obtenir des résultats pour différents domaines de la population (Hogan 2003). Le but du projet n'est pas d'ajuster les chiffres du recensement mais d'obtenir des informations sur la qualité du recensement de l'an 2000 et les potentiels d'améliorations pour les recensements suivants.

Cet article présente les différentes étapes des estimations et les résultats. Les sections 2 et 3 décrivent les jeux de données et les estimateurs de couverture. La section 4 expose les détails de la construction des différents status utilisés dans les estimateurs. La section 5 décrit l'approche utilisée pour comparer les valeurs relevées dans le recensement et dans l'enquête pour les personnes appariées de S_p . Les sections 6 et 7 présentent les résultats numériques et la conclusion.

rapport aux hypothèses généralisées des modèles linéaires mixtes. Cette conclusion peut revêtir une importance particulière pour toute application de la théorie des modèles linéaires mixtes fondés sur la normalité aux ensembles de données où les hypothèses de normalité ne s'avèrent pas tout à fait adéquates, comme c'est le cas des données sur le revenu.

La recherche a été financée en partie par Miur-PRIN 2003 *Statistical analysis of changes of the Italian productive sectors and their territorial structure*, sous la coordination du professeur C. Filippucci. Les travaux d'Emilio Fabrizi ont été subventionnés en partie par l'université de Bergamo (subventions n° 60FABR06 et 60BIFP04).

Nous remercions ISTAT d'avoir eu la gentillesse de nous fournir les données dont nous nous sommes servis pour réaliser nos travaux.

Bibliographie

- Battese, G.E., Harter, R.M., et Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.
- Bell, W. (2001). Discussion with "Jackknife in the Fay-Herriot model with an example". *Proceeding of the Seminar on Funding Opportunity in Survey Research*, 98-104.
- Betti, G., et Verma, V. (2002). Non-monetary or lifestyle deprivation, in EUROSTAT (2002). *Income, Poverty Risk and Social Exclusion in the European Union*, Second European Social Report, 87-106.
- Butar, F., et Lahiri, P. (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112, 63-76.
- Datta, G.S., et Lahiri, P. (1999). A unified measure of uncertainty of estimated best linear predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Datta, G.S., Lahiri, P., et Maiti, T. (2002). Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.
- Datta, G.S., Lahiri, P., Maiti, T., et Lu, K.L. (1999). Hierarchical bayes estimation of unemployment rates for the States of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- EURARARE CONSONOTUM (2004). EURAREA. Enhancing Small Area Estimation Techniques to meet European Needs, Project Reference Volume, <http://www.statistics.gov.uk/cwurare/download.asp>.
- EUROSTAT (2002). European social statistics - Income, poverty and social exclusion. 2^{ème} rapport.
- Falorsi, P.D., Falorsi, S., et Russo, A. (1994). Comparison empirique de méthodes d'estimation pour petites régions pour l'enquête sur la population active italienne. *Techniques d'enquête*, 20, 179-184.
- Falorsi, P.D., Falorsi, S., et Russo, A. (1999). Small area estimation at provincial level in the Italian Labour Force Survey. *Journal of the Italian Statistical Society*, 1, 93-109.
- Gejman, A., Lehtonen, R., et Särndal, C.-E. (2005). The Effect of Model Quality on Model-Assisted and Model-Dependent Estimators of Totals and Class Frequency for Domains, Article présenté à la Conférence de SAFE2005, Challenges in Statistics Production for Domains and Small Areas, 28-31 août 2005, Jyväskylä, Finland.
- Verbeke, G., et Molenberg, H. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.
- Ghosh, M., Nangia, N., et Kim, D. (1996). Estimation of median income of four-person families: A bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Heady, P., Higgins, N., et Ralphs, M. (2004). Evidence-Based Guidance on the Applicability of Small Area Estimation Techniques. Article présenté à l'European Conference on Quality and Methodology in Official Statistics, Mainz, Allemagne, 24-26 mai.
- Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics*, 24, 255-286.
- Jiang, J., Lahiri, P., et Wan, S.M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.
- Jiang, J., et Lahiri, P. (2006a). Estimation of finite population domain means - A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.
- Jiang, J., et Lahiri, P. (2006b). Mixed model prediction and small area estimation. Discussion sollicitée par l'éditeur, *Test*, 15, 1-96.
- Kenward, M.G., et Roger, J.H. (1986). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983-997.
- Lehtonen R., Särndal C.-E. et Viitanen A. (2003) L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 1 37-49.
- Pfeiffermann, D. (2002). Small area estimation - New developments and directions. *Revue internationale de Statistique*, 70, 125-143.
- Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.
- Rao, J.N.K. (2003). *Small Area Estimation*, New York: John Wiley & Sons, Inc.
- Rao, J.N.K., et Yeu, M. (1994). Small area estimation by combining time series and cross-sectional data. *Canadian Journal of Statistics*, 22, 511-528.
- Rendtel, U., Behr, A. et Sisto, J. (2003). Attribution effect in the European Community household panel, CHINTX PROJECT, European Commission.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Singh, A.C., Mantel, H.J. et Thomas, B.W. (1994). MPLSE à données chronologiques pour petites régions évalués à l'aide de données d'enquête. *Techniques d'enquête*, 20, 35-46.
- Sjöqvist, E., et Thomsen, I. (1987). Application of some empirical bayes methods to small area statistics. *Bulletin of the International Statistical Institute*, 4, 435-450.
- Vandecasteele, L., et Debeels, A. (2004). Modelling attrition in the European Community Household Panel: The effectiveness of weighting. 2^{ème} Conférence Internationale des utilisateurs de l'ECHP, EPUNet 2004, Berlin, juin 24-26.
- Viitanen, R., et Särndal, C.-E. (2005). The Effect of Model Quality on Model-Assisted and Model-Dependent Estimators of Totals and Class Frequency for Domains, Article présenté à la Conférence de SAFE2005, Challenges in Statistics Production for Domains and Small Areas, 28-31 août 2005, Jyväskylä, Finland.
- Verbeke, G., et Molenberg, H. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

l'estimateur Prasad-Rao (14), les résultats de l'énoncé ESTIMATE de Proc MIXED sont utilisés et l'option $g_2(\psi)$ est obtenue des résultats de Proc MIXED. L'option $g_2(\psi)$ est obtenue des résultats de Proc MIXED. L'option KENWARDROGER est activée. La somme $g_1(\psi) + g_2(\psi)$ est obtenue des résultats de Proc MIXED. L'option KENWARDROGER permet le calcul du facteur d'inflation de l'EQM décrit dans Kenward et Rogers (1986), lequel est équivalent à $2g_3(\psi)$ (voir aussi Rao 2003, section 6.2.7). L'estimateur $eqm_{BL}^{(n_{PT}^{MPLSBE})}$ est fondé sur le ré-échantillonnage. Donc, l'évaluation de son rendement dans le cadre d'un exercice de Monte Carlo nécessite la mise en oeuvre de deux simulations emboîtées : pour chaque r ($r = 1, \dots, R$), nous exécutons les répétitions R_{BOOT} nécessaires pour estimer les attentes à l'égard du modèle bootstrap. Pour atténuer le fardeau de calcul, nous établissons que $R_{BOOT} = 150$. Butar et Lahiri (2003) proposent une approximation analytique de l' eqm_{BL} , mais uniquement dans le cas des modèles qui ne sont pas aussi complexes que celui qui nous occupe.

Aussi bien pour $eqm_{BL}^{(n_{PT}^{MPLSBE})}$ que pour $eqm_{JLW}^{(n_{PT}^{MPLSBE})}$, nous avons préparé des codes SAS spéciaux utilisant les résultats de Proc MIXED.

Pour comparer les trois estimateurs de l'EQM, nous employons les mêmes mesures qui sont utilisées pour évaluer le rendement des estimateurs ponctuels, BRAM et MEQMR. Compte tenu du risque de sous-estimation des estimateurs de l'EQM, nous nous intéressons également à tout indice de biais associé aux estimateurs en question. Par conséquent, dans le cas des estimateurs de l'EQM, en plus de la moyenne des valeurs absolues des estimations obtenues pour le biais de chaque région (BRAM), nous calculons également la moyenne de ces estimations sans tenir compte de la valeur absolue (BRAM'), de manière à mieux comprendre si les estimateurs donnés ont effectivement ment tendance à sous-évaluer l'EQM. Donc, les mesures calculées s'énoncent comme suit :

$$BRAM = m^{-1} \sum_{r=1}^m \left\{ R^{-1} \sum_{j=1}^R \left(eqm_{ACT}^{(n_{PT}^{MPLSBE})} - 1 \right) \right\}$$

$$BRAM' = m^{-1} \sum_{r=1}^m \left\{ R^{-1} \sum_{j=1}^R \left(eqm_{ACT}^{(n_{PT}^{MPLSBE})} - 1 \right) \right\}$$

$$MEQMR = m^{-1} \sum_{r=1}^m \left\{ R^{-1} \sum_{j=1}^R \left(eqm_{ACT}^{(n_{PT}^{MPLSBE})} - 1 \right) \right\}$$

où le symbole* renvoie aux procédures d'estimation envisagées, soit PR, BL et JLW. Les résultats des comparaisons en fonction des $R = 500$ itérations de MC sont indiqués dans le tableau 3.

En ce qui concerne MEQMR et BRAM, les trois estimateurs se comportent de façon semblable, et aucun d'entre eux ne se démarque nettement des deux autres. Néanmoins, la colonne BRAM' démontre clairement que

Pour ce qui est des résultats du rééchantillonnage, on peut aussi les considérer acceptables dans ce cas-cl, et on constate un lien entre les résultats obtenus pour les estimateurs MPLSBE dérivés de modèles analogues avec ou sans covariables.

4.2 Comparaison de différents estimateurs de l'EQM associée aux estimateurs MPLSBE

À la section 2.3, nous avons examiné trois types d'estimateurs de l'EQM associés aux estimateurs MPLSBE. Dans la présente section, nous allons comparer le rendement de ces trois estimateurs au moyen d'un exercice de simulation. Comme nous attacherons sur une estimation de l'EQM plutôt qu'une comparaison des estimateurs MPLSBE dérivés de différents modèles, nous tiendrons seulement compte du prédicteur associé au modèle MM_5 , dont le rendement a été jugé supérieur aux autres à la section précédente.

Supposons que nous désignons le prédicteur de \hat{Y}_{PT} par \hat{Y}_{PT}^{MPLSBE} et son erreur quadratique moyenne par $EQM(\hat{Y}_{PT}^{MPLSBE})$. L'estimateur suivant :

$$eqm_{ACT}^{(n_{PT}^{MPLSBE})} = \frac{1}{R} \sum_{r=1}^R \left(\hat{Y}_{PT}^{MPLSBE} - \hat{Y}_{PT}^{ACT} \right)^2 + \left(\hat{Y}_{PT}^{MPLSBE} - \hat{Y}_{PT}^{ACT} \right)^2$$

où \hat{Y}_{PT}^{MPLSBE} est \hat{Y}_{PT}^{MPLSBE} calculé en fonction du r^e échantillon répété et $\hat{Y}_{PT}^{ACT} = R^{-1} \sum_{r=1}^R \hat{Y}_{PT}^{ACT}$, servira de point de repère pour comparer le rendement des estimateurs de l'erreur quadratique moyenne décrits à la section 2.3, parce que l'erreur quadratique moyenne réelle est inconnue. Comme dans le cas des estimateurs ponctuels, tous les calculs sont effectués au moyen de SAS. Pour déterminer

Ce dernier résultat est vraiment encourageant. En fait, dans le contexte de l'estimation pour petits domaines, l'absence des totaux connus des covariables de la population peut s'avérer fort contraignante quand vient le temps d'obtenir des estimations fiables. La réduction de MEQMR observée relativement à la considération de plusieurs cycles d'une enquête par panel révèle que les estimations peuvent être améliorées grâce à l'emprunt d'information au fil du temps, lorsqu'il n'est pas possible de mettre à profit l'information auxiliaire.

Pour ce qui est des résultats du rééchantillonnage, on peut aussi les considérer acceptables dans ce cas-cl, et on constate un lien entre les résultats obtenus pour les estimateurs MPLSBE dérivés de modèles analogues avec ou sans covariables.

Pour ce qui est du biais, les estimateurs MPLSBE déclinant des modèles sans covariables ont tendance à être plus biaisés que les modèles correspondants comportant des covariables.

L'analyse de la colonne $MBFF_{trans}$ révèle que la réduction de MEQMR permise par certains des modèles empruntant de l'information au fil du temps est plus prononcée lorsque que des covariables sont incluses; en effet, elle atteint 22 % dans le meilleur exemple du modèle MM_5 .

En ce qui concerne le biais, l'estimateur GREG donne la valeur de BRAM semblable. Parmi les estimateurs MPLSBE, ceux qui sont associés aux modèles *MM1* et *MM5* sont plus efficaces en ce qui concerne MEQMR, mais ils sont généralement plus biaisés que l'estimateur du *SMM*. Cette situation est probablement attribuable au fait que notre évaluation du rendement se limite au dernier cycle; pour ce sous-ensemble de données, nous nous attendons à ce que l'ajustement de la régression sous-jacente de *SMM*, en fonction du dernier cycle seulement, soit meilleur que l'ajustement fondé sur l'ensemble des données. Pour ce qui est des estimateurs MPLSBE, la colonne MEFF^{Trans} indique comment les gains d'efficacité des prédicteurs, en fonction de l'emprunt d'information au fil du temps, s'avèrent parfois positifs, parfois négatifs. Les modèles *MM2* et *MM4* (voir les formules (6) et (8)), où les effets de l'interaction entre la région et la période sont présents, sont apparemment inadéquats, étant donné que les prédicteurs associés aux deux modèles donnent des résultats plutôt médiocres. Le rendement du prédicteur associé à *MM3* (voir (7)) est légèrement inférieur à celui du prédicteur associé au modèle transversal. Ce résultat plutôt surprenant est probablement attribuable au petit nombre de cycles, qui ne permet pas d'estimer efficacement le coefficient de corrélation entre les divers effets temporels consécutifs. Comme nous l'avons déjà mentionné, l'estimateur associé au modèle *MM5* est celui qui offre le meilleur rendement : il est considérablement plus efficace que celui de *SMM*, et affiche un ratio MEFF^{Trans} d'environ 85 %, ce qui représente un gain d'efficacité d'environ 15 % si l'on tient compte de plus d'un cycle. L'estimateur MPLSBE associé à *MM1* se révèle également plus efficace que celui de *SMM*, mais dans ce cas-ci, le gain d'efficacité s'établit à seulement 5 %. Ces résultats confirment le fait que les données au niveau du ménage à plusieurs moments consécutifs peuvent être utilisées, au moyen de certains types de modèles longitudinaux, pour produire des estimations plus efficaces.

Si nous passons à l'indicateur de rétrécissement mentionné à la dernière colonne du tableau, nous voyons que l'estimateur direct sursumme l'écart-type de la population des moyennes régionales de 15 %. On observe le même effet, quoique quelque peu atténué, pour ce qui est de l'estimateur GREG, dont l'écart-type est gonflé de 10 %. En revanche, l'estimateur COMP a tendance à réduire les estimations vers le centre de la répartition, ce qui entraîne une réduction de l'écart-type des moyennes régionales d'environ 10 % par rapport à la population. Ces résultats s'alignent à ceux qu'on obtient d'autres auteurs qui ont comparé les mêmes types d'estimateurs (Heady et coll. 2004; Spjøtvoll et Thomson 1987). Les résultats obtenus pour les estimateurs des MPLSBE sont plus encourageants, étant donné que la différence calculée en pourcentage est toujours inférieure à 10 % en chiffres absolus. Ainsi, à cet égard, tous les estimateurs MPLSBE semblent acceptables. De plus, nous pouvons nous attendre à ce que les estimateurs MPLSB soient sous-dispersés comparativement aux paramètres de population correspondants. En pareil cas, l'indicateur ETER prend des valeurs positives pour certains estimateurs longitudinaux MPLSBE, parce qu'il est calculé seulement en fonction du dernier cycle, tandis que les modèles longitudinaux visent à prédire $m \times T$ paramètres. Le tableau 2 résume les résultats au sujet des estimateurs MPLSBE qui sont associés aux modèles de variance de l'erreur aléatoire, comme décrit au dernier paragraphe de la section 2. Lorsque les modèles ne comportent aucune variable auxiliaire, l'avantage de l'emprunt d'information selon la période et la région ressort indépendamment de l'avantage associé aux covariables. Comme prévu, les gains d'efficacité mesurés par MEFF^{Dir} sont inférieurs à ceux du tableau 1, bien que les réductions de MEQMR demeurent importantes. Le classement des prédicteurs associés à la spécification des divers effets aléatoires demeure le même que celui du tableau 1, le prédicteur associé au modèle *MM5* s'avérant le plus efficace, comme l'indique la colonne MEQMR%. Les gains d'efficacité associés à ce dernier estimateur par rapport à l'estimateur direct se chiffrent à environ 43 %.

Tableau 2 Indicateurs de rendement – aucune information auxiliaire disponible

Modèle	BRAM%	MEQMR%	MEFF ^{Dir} %	MEFF ^{Trans} %	ETER%
<i>SMM</i>	2,7	0,575	72,8	100,0	-7,6
<i>MM1</i>	2,9	0,556	70,3	96,6	7,5
<i>MM2</i>	2,8	0,639	80,8	111,0	-3,0
<i>MM3</i>	3,7	0,574	72,6	99,7	8,6
<i>MM4</i>	3,5	0,691	87,2	119,8	-6,7
<i>MM5</i>	3,0	0,445	56,2	77,2	-6,3

où \bar{Y}_T correspond aux valeurs de la moyenne de population des régions m pendant la période T , avec l'écart-type empirique des valeurs régionales estimées, qui sont obtenues par la formule qui suit dans le cadre d'une étude par simulation :

$$etc = R^{-1} \sum_{R=1}^R \left[\sqrt{m^{-1} \sum_{m=1}^m (\bar{Y}^{JT(r)} - \bar{Y}^{T(r)})^2} \right]$$

où $\bar{Y}^{T(r)}$ est la moyenne des valeurs estimées pour les m régions pendant la période T dans le cadre de l'exécution de la simulation r . La comparaison s'effectue au moyen de l'indicateur

$$ETER = \frac{ETE}{ete} - 1 \tag{19}$$

qui révèle en quoi l'écart-type empirique associé à un estimateur en particulier diffère de celui de la population.

Le tableau 1 indique les valeurs en pourcentage de BRAM, MEQMR, MEFF et ETER obtenues pour l'estimateur direct, les estimateurs fondés sur le plan de sondage selon les formules (2) et (3) et les estimateurs des MPLSBE dérivés à partir des modèles (5) - (10).

Tous les estimateurs fonctionnent considérablement mieux que $\bar{Y}^{T,DIR}$ en ce qui concerne MEQMR et donnent des valeurs MEFF^{Dir} de moins de 100 %. Nous constatons également que les estimateurs fondés sur le plan de sondage donnent de moins bons résultats que les estimateurs MPLSBE pour ce qui est de MEQMR, et que les gains d'efficacité démontées par MEFF^{Dir} sont particulièrement élevées dans certains cas (plus de 50 %). Ce résultat souligne l'exactitude supérieure des estimateurs fondés sur le modèle en question.

L'estimateur MPLSBE le plus fiable est celui du modèle $MM5$, où les effets régionaux et temporels sont indépendants et les résidus sont autocorrélés conformément à un processus AR(1), ce qui entraîne des gains d'efficacité d'environ 60 % par rapport à l'estimateur direct. Vient ensuite l'estimateur MPLSBE associé au modèle $MM1$, qui diffère du premier estimateur en raison de l'absence de résidus autocorrélés.

Modèle	BRAM%	MEQMR%	MEFF ^{Dir} %	MEFF ^{Trans} %	ETER%
DIR	0,0	0,787	100,0	-	15,6
COMP	2,7	0,552	70,1	-	-9,8
GRGC	0,2	0,543	68,2	-	10,0
SMM	2,3	0,377	47,7	100,0	-8,7
MM1	3,1	0,358	45,3	95,0	2,4
MM2	2,4	0,427	54,1	113,4	-4,4
MM3	2,6	0,380	48,3	101,2	4,7
MM4	2,6	0,429	54,2	113,6	-8,0
MM5	2,9	0,318	40,4	84,7	-7,0

Tableau 1 Indicateurs de rendement - information auxiliaire disponible

$$ETE = \sqrt{m^{-1} \sum_{m=1}^m (Y^{JT} - \bar{Y}^T)^2}$$

Le ratio est désigné MEFF^{Trans}. En ce qui concerne l'évaluation du degré de réticissément, nous avons comparé l'écart-type empirique des valeurs de la population du domaine :

où $\bar{Y}^{JT(r)}$ représente l'estimation pour la région d , la période T et l'échantillon répété r , tandis que \bar{Y}^{JT} est la moyenne de la population estimée. Soulignons que BRAM mesure le biais d'un estimateur, alors que MEQMR en mesure la précision. Le nombre de répétitions de R est établi à 500, ce qui devrait suffire pour obtenir les estimations Monte Carlo stables des valeurs prévues et des variances, une méthode souvent utilisée dans les études de simulation au sujet de l'estimation pour petits domaines (Headly, Higgins et Ralphs 2004; EURAREA Consortium 2004).

Pour évaluer le rendement des estimateurs, nous avons adopté une approche souvent mentionnée dans la documentation (voir Rao 2003; section 7.2), au moyen de deux indicateurs : le biais relatif absolu moyen (BRAM) et la moyenne des erreurs quadratiques moyennes relatives (MEQMR) :

L'asymétrie apparente des résidus suggère également que l'hypothèse de normalité pour les erreurs ne s'avère pas tout à fait adéquate. Nous maintenons cette hypothèse pour tous les modèles que nous précisons, et nous servons des estimateurs MVR pour les composantes de variance. En fait, nous pouvons nous attendre à ce que les écarts de la normalité aient un léger effet sur les valeurs ponctuelles des prédicteurs. Les formules des MPLSB peuvent être dérivées sans qu'il n'y ait normalité, qu'il plus est, nous avons de très bonnes raisons de nous attendre à ce que les estimateurs MVR (et MV) de ψ donnent de bons résultats, même si la normalité ne tient pas le coup (pour plus de détails, voir Jiang 1996). Les écarts de la normalité peuvent avoir des conséquences plus graves sur l'estimation de l'EQM, et nous nous pencherons sur ce problème à la section 4.2 ci-dessous.

4. Résultats

4.1 Les estimateurs ponctuels

Tous les calculs nécessaires dans l'exercice de simulation décrit à la section 3 ont été effectués au moyen de la version 9.1 de SAS pour Windows. Les estimateurs MPLSB sont obtenus à l'aide de Proc MIXED, et la création des échantillons est fondée sur Proc SURVEYSELECT.

Étant donné que l'estimation précise des paramètres régionaux est l'objectif principal de l'estimation pour petits domaines, nous avons d'abord évalué le rendement des estimateurs décrits pour prédire les valeurs des régions individuelles. Nous avons aussi évalué l'ampleur du surestimationnement rattaché à chaque estimateur. En fait, les estimations pour petits domaines devraient refléter (au moins de façon approximative) la variabilité globale des paramètres régionaux sous-jacents.

Mentionnons que notre essai de simulation a pour objet d'évaluer les propriétés fondées sur le plan de sondage des estimateurs, c'est-à-dire que les échantillons aléatoires sont générés à partir d'une population fixe.

Les manuels académiques (voir la section 2.2), les manuels d'enquête (voir Verbeke et Molenberghs 2000, chapitre 9) recommandent souvent la méthode suivante, qui a été suivie dans ce cas-ci : d'abord, nous appliquons à nos données la méthode standard de régression des moindres carrés ordinaires en utilisant toutes les covariables disponibles; ensuite, nous analysons les résidus obtenus pour nous aider à repérer les effets aléatoires. Cette analyse préliminaire a été effectuée séparément sur plusieurs échantillons aléatoires de 1000 unités, établis conformément au plan de rééchantillonnage susmentionné.

Le R^2 rajusté de la régression des moindres carrés ordinaires se rapproche de 0,35 dans chacun des échantillons observés. Ce résultat plutôt faible est attribuable à la nature du phénomène à l'étude (le revenu du ménage n'est pas facile à prédire), à l'information contenue dans l'enquête, et à la contrainte de la nécessité d'inclure seulement les covariables pour lesquelles le chiffre de population peut être obtenu du recensement.

La figure 1 contient des diagrammes des quartiles des résidus par région et par cycle, créés pour un des échantillons de Monte Carlo (les constatations sont très semblables d'un échantillon à l'autre). L'analyse des diagrammes suggère une corrélation intrarégionale et à l'intérieur des cycles, et donc la nécessité de préciser les modèles, comprenant les effets des régions et des cycles. Lorsque l'on analyse les résidus, les avantages de l'inclusion des effets intracarrés (c'est-à-dire les effets régionaux fluctuant au fil du temps) sont moins clairs.

Par ailleurs, les résidus affichent une certaine autocorrélation, la moyenne du coefficient d'autocorrélation calculée pour tout les historiques individuels de valeurs résiduelles se chiffrant à 0,27. Bien que ce niveau d'autocorrélation ne soit pas très élevé, par souci d'intégrité, nous avons décidé de tenir également compte des modèles comportant des erreurs autocorrélées ou des effets aléatoires. Après avoir mis à l'essai diverses structures d'autocorrélation (ARMA(p , q), modèle linéaire généralisé, etc.), nous avons constaté que le processus autorégressif de l'ordre 1 est le mieux adapté à nos données.

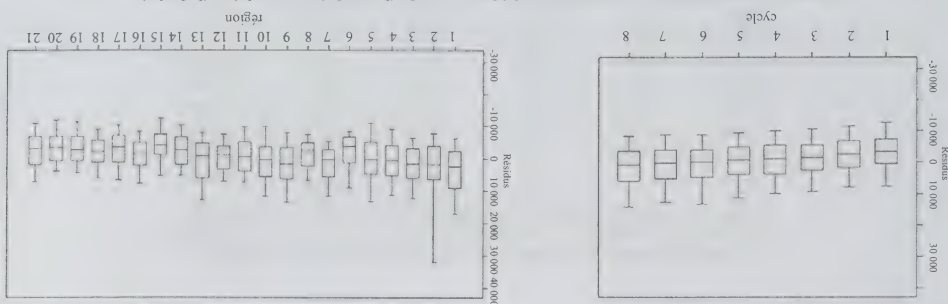


Figure 1 Diagramme des quartiles des résidus par cycle (à gauche) et par région (à droite)

empirique du rendement de ces trois estimations dans un contexte où le nombre de régions est modéré et l'hypothèse de normalité pourrait ne pas s'avérer tout à fait adéquate. Voici une brève description des trois méthodes d'estimation. Soit la formule $E\mathbb{Q}M[\eta_{MPLSBE}(\psi) - \eta_{MPLSBE}(\psi) - n]^2$, où l'espérance renvoie au modèle (4). Il est possible de démontrer que, dans l'optique de la normalité,

$$E\mathbb{Q}M[\eta_{MPLSBE}(\psi)] = g_1(\psi) + g_2(\psi) + E(\eta_{MPLSBE} - \eta_{MPLSBE})^2 \quad (12)$$

où $g_1(\psi) = \mathbf{k}(\mathbf{G} - \mathbf{G}\mathbf{Z}\mathbf{V}^{-1}\mathbf{Z}\mathbf{G})\mathbf{k}'$ et $g_2(\psi) = \mathbf{d}'(\mathbf{X}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}$, et $\mathbf{d} = \mathbf{m}' - \mathbf{k}\mathbf{G}\mathbf{Z}\mathbf{V}^{-1}\mathbf{X}$ (voir Rao 2003, chapitre 3). Au moyen de l'approximation suivante, fondée sur un argument de la série de Taylor

$$E(\eta_{MPLSBE} - \eta_{MPLSBE})^2 \approx \text{tr}[(\partial\mathbf{b}'/\partial\psi)(\mathbf{V}(\partial\mathbf{b}'/\partial\psi))'\mathbf{V}(\psi)] = g_3(\psi) \quad (13)$$

Souignons qu'ici, \approx signifie que les termes omis sont de l'ordre de $o(m^{-1})$. On obtient un estimateur asymptotiquement sans biais de (13), selon Prasad et Rao (1990), grâce à la formule suivante :

$$\text{eqm}^{pr}(\eta_{MPLSBE}) = g_1(\psi) + g_2(\psi) + 2g_3(\psi). \quad (14)$$

Datta et Lahiri (1999) démontrent que, en supposant la normalité avec l'estimation de ψ , au moyen des méthodes MVR ou MV, $\text{eqm}^{pr}(\eta_{MPLSBE})$ estime $E\mathbb{Q}M[\eta_{MPLSBE}(\psi)]$ avec un biais de l'ordre $o(m^{-1})$.

Blair et Lahiri (2003) proposent une estimation paramétrique bootstrap selon le modèle (13), en s'appuyant sur une hypothèse de normalité. Nous adaptons leur estimateur aux modèles que nous analysons, en présument de l'utilisation du modèle bootstrap suivant :

$$y^* | \mathbf{v}^* \sim N[\mathbf{X}\beta + \mathbf{Z}\mathbf{v}^*, \mathbf{R}(\psi)] \quad (1)$$

$$(15)$$

$$\mathbf{v}^* \sim N[\mathbf{0}, \mathbf{G}(\psi)] \quad (11)$$

où $\mathbf{v} = (v_1, \dots, v_J, v_J^*)'$. La méthode du bootstrap paramétrique est ensuite utilisée à deux reprises, une fois pour estimer les deux premiers termes de (13), ce qui permet de corriger les biais de $g_1(\psi) + g_2(\psi)$, et une fois pour estimer $g_3(\psi)$. L'estimateur de (13) suivant est proposé :

$$\text{eqm}^{bl}(\eta_{MPLSBE}) = 2[g_1(\psi) + g_2(\psi)] - E[g_1(\psi^*) + g_2(\psi^*)] + E[g_1(\eta(\mathbf{y}, \beta(\psi)), \psi^*) + g_2(\eta(\mathbf{y}, \beta(\psi)), \psi^*)] \quad (16)$$

pondération.

où ψ_j^* est l'estimation de ψ calculée au moyen de toutes les données sauf celles de la région j . Parallèlement, $\eta_{MPLSBE}^j = \eta_{MPLSBE}(\mathbf{y}^j, \beta(\psi_j^*), \psi_j^*)$. Il est important de souligner que, si l'on se fie aux résultats de la simulation signalés dans Jiang et coll. (2002), l' eqm^{JLW} est considéré plus robuste que l' eqm^{pr} , en ce qui a trait aux écarts par rapport à l'hypothèse de normalité, qu'risque également d'être essentielle à l'estimateur eqm^{bl} .

3. L'étude par simulation fondée sur les données du panel communautaire de ménages

La population cible du PCM est composée de tous les ménages d'un vaste sous-ensemble de pays membres de l'UE. Bien que les lignes directrices générales de l'enquête aient été instaurées par Eurostat, une certaine souplesse a été permise, de sorte que le plan d'échantillonnage comporte des différences d'un pays à l'autre. Pour ce qui est de l'Italie, l'enquête est fondée sur un plan stratifié à deux degrés, selon lequel les strates ont été formées par le regroupement des UPE (municipalités) en fonction de la région géographique (NUTS2) et de la taille de la population. Le PCM traite la non-réponse des unités, l'érosion de l'échantillon et les nouvelles entrées au moyen de la pondération et de l'imputation. Étant donné que l'érosion pourrait donner lieu à des estimations biaisées du revenu si elle n'apparaît pas de façon aléatoire, l'effet de la pauvreté sur la propension au décrochage a été étudié (Rendic, Behr et Sisto 2003; Vandecasteele et Debeis 2004), et les résultats de ces études révèlent que dans certains pays, notamment en Italie, cet effet disparaît sous le contrôle des variables de

Le troisième modèle :

MM3 : $y_{dt} = \mathbf{x}_{dt}^* \beta + v_d + \alpha_i^* + e_{dt}$ (7)

est obtenu en supposant que $s = 2, q_1 = m, q_2 = T, \mathbf{G}_1 = \sigma_y^2 \mathbf{I}_m, \mathbf{R} = \sigma_y^2 \mathbf{I}_n$, tandis que l'élément générique $g_2(h, k)$ de \mathbf{G}_2 est $g_2(h, k) = \sigma_y^2 \rho_{h-k}^{\alpha}$, $h, k = 1, \dots, T$. Certains effets régionaux et temporels sont indépendants les uns des autres, tout comme c'est le cas du modèle MM1, mais nous présumons que les effets temporels suivent un processus AR(1).

Le quatrième modèle :

MM4 : $y_{dt} = \mathbf{x}_{dt}^* \beta + \delta_{dt}^* + e_{dt}$ (8)

est semblable au modèle MM2 en ce qu'il est caractérisé par des effets régionaux variant au fil du temps, mais nous présumons en outre que ces effets suivent un processus AR(1). Donc, à condition d'ordonner les observations par région, en ce qui concerne la formule générale (4), nous obtenons $s = 1, q_1 = mq, \mathbf{G}_1 = \text{diag}(\mathbf{G}_1^q), \mathbf{R} = \sigma_y^2 \mathbf{I}_n$ où \mathbf{G}_1^q , $d = 1, \dots, m$, est une matrice $T \times T$ dont l'élément générique $g_1^q(h, k) = \sigma_y^2 \rho_{h-k}^{\alpha}$, $h, k = 1, \dots, T$.

La dernière spécification :

MM5 : $y_{dt} = \mathbf{x}_{dt}^* \beta + v_d + \alpha_i + e_{dt}^*$ (9)

peut être obtenue au moyen de (4) avec $s = 2, q_1 = m, q_2 = T, \mathbf{G}_1 = \sigma_y^2 \mathbf{I}_m, \mathbf{G}_2 = \sigma_y^2 \mathbf{I}_T$. À condition d'ordonner les observations selon le ménage et la période $\mathbf{R} = \text{diag}(\mathbf{R}^{dt})$ où \mathbf{R}^{dt} est une matrice $T \times T$ dont l'élément générique est donné par $r_{dt}^{\alpha}(h, k) = \sigma_y^2 \rho_{h-k}^{\alpha}$, $h, k = 1, \dots, T$. Comme dans le modèle MM1, il existe des effets régionaux et temporels indépendants les uns des autres, mais nous supposons que les erreurs sont autocorrélées selon un processus AR(1).

Pour évaluer les conséquences de l'utilisation des cycles précédents de l'enquête sur l'efficacité de l'estimateur, le modèle linéaire transversal mixte (SMM) fondé exclusivement sur les données du dernier cycle T a servi de point de repère :

SMM : $y_{dt} = \mathbf{x}_{dt}^* \beta + \theta_d + e_{dt}$ (10)

où $\theta_d \sim N(0, \sigma_\theta^2)$, $e_{dt} \sim N(0, \sigma_e^2)$. Encore une fois, il s'agit d'un cas particulier de (4) obtenu pour $s = 2, q_1 = m, \mathbf{G}_1 = \sigma_y^2 \mathbf{I}_m$ et $\mathbf{R} = \sigma_y^2 \mathbf{I}_n$. Soulignons que (10) est le modèle standard de régression à erreurs emboîtées de Battese et coll. (1988).

Nous tenons également compte des modèles linéaires de variance de l'erreur aléatoire (voir Rao 2003; section 5.5.2) obtenus en remplaçant $\mathbf{x}_{dt}^* \beta$ dans les formules (5) - (10) par un point d'ordonnée à l'origine général de θ . Ces modèles seront désignés comme suit : MM1*, MM2*, MM3*, MM4*, MM5*, SMM*. Toutes les hypothèses au

sujet des effets aléatoires et des résidus demeurent inchangées. Ce dernier groupe de modèles nous permet d'explorer les gains d'efficacité découlant de la mise à profit de la répétition de l'observation de la même unité lorsqu'il n'existe pas de covariables au niveau de population. En ce qui concerne l'estimation pour petits domaines, l'objectif est de prédire les combinaisons scalaires linéaires des effets fixes et aléatoires du type $\eta = \mathbf{m}'\beta + \mathbf{k}'v$ où \mathbf{m} et \mathbf{k} renvoient respectivement aux vecteurs $d \times 1$ et $q \times 1$, et où $q = \sum_j q_j$. Nous obtenons le meilleur prédicteur linéaire sans biais (MPLSB) de η en estimant (β, v) pour réduire l'EQM du modèle dans tous les estimateurs linéaires :

$\eta_{\text{MPLSB}}(\psi) = \mathbf{m}'\beta(\psi) + \mathbf{k}'v(\psi)$ (11)

Lorsque les composantes de la variance de ψ sont inconnues, nous pouvons les estimer à partir des données et les substituer dans la formule (11), en vue d'obtenir les «MPLSB empiriques» (ou MPLSBE) $\hat{\eta}_{\text{MPLSBE}}(\psi) = \mathbf{m}'\hat{\beta}(\psi) + \mathbf{k}'\hat{v}(\psi)$ (pour plus de détails, voir Rao 2003, chapitre 6, et Jiang et Lahiri 2006b).

En ce qui a trait à l'estimation de ψ , plusieurs méthodes ont été proposées dans la documentation, telles que le maximum de vraisemblance (MV), et le maximum de vraisemblance restreint (MVR), qui présument de la normalité des termes aléatoires, et la méthode MINQUE proposée par Rao (1972), qui est non paramétrique. Pour les besoins du présent document, nous avons opté pour la méthode MVR, ce qui signifie donc que nous avons supposé la normalité.

2.3 Les mesures de l'incertitude associée aux prédicteurs fondés sur les modèles linéaires mixtes

Dans la documentation pour petits domaines, diverses approches sont proposées pour résoudre la problématique de l'estimation de l'EQM des MPLSBE, en tenant compte de la variabilité des composantes estimées de la variance et de la covariance.

Une méthode populaire est fondée sur une approximation de l'EQM par la série de Taylor sous l'hypothèse de la normalité (Prasad et Rao 1990; Datta et Lahiri 1999). Plus récemment, grâce aux ordinateurs à haute vitesse et aux logiciels puissants, des méthodes de rééchantillonnage ont été proposées. Par exemple, Butar et Lahiri (2003) présentent une méthode « bootstrap » paramétrique qui est fondée sur l'hypothèse de normalité, mais moins lourde du point de vue de l'analyse que la méthode de la série de Taylor. Jiang et coll. (2002) décrivent une méthode jackknife générale, qui nécessite une hypothèse de répartition plus faible que la normalité (linéarité postérieure). Notre objectif est d'effectuer une comparaison

plusieurs avantages, dont le plus important est la possibilité de mettre à l'essai les hypothèses sous-jacentes. Les modèles linéaires mixtes sont très souvent utilisés pour estimer les moyennes ou les totaux des variables continues sur des petits domaines. Un modèle linéaire mixte général peut se décrire comme suit :

$$(4) \quad \mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{v}_1 + \dots + \mathbf{Z}_s\mathbf{v}_s + \mathbf{e},$$

où $\mathbf{y} = \{y_{dh}\}$ est le vecteur de dimension n des observations des échantillons, β est un vecteur $p \times 1$ des effets fixes, \mathbf{v}_j est un vecteur $q_j \times 1$ des effets aléatoires ($j = 1, \dots, s$), $\mathbf{e} = \{e_{dh}\}$ est un vecteur de erreurs; nous présumons que \mathbf{X} est de rang p , $\mathbf{Z}_j' = \{\mathbf{z}_{dhj}'\}$ est une matrice $n \times q_j$ de l'incidence de l'effet aléatoire j^e . Nous supposons que $E(\mathbf{v}_j) = 0$, $V(\mathbf{v}_j) = \mathbf{G}_j$, $E(\mathbf{e}) = 0$, $V(\mathbf{e}) = \mathbf{R}$ (toutes les espérances sont par rapport au modèle (4)) et que $\mathbf{v}_1, \dots, \mathbf{v}_s, \mathbf{e}$ sont mutuellement indépendants. Ainsi, la matrice de variance-covariance \mathbf{y} se calcule comme suit :

$$\mathbf{V} = V(\mathbf{y}) = \sum_{j=1}^s \mathbf{Z}_j \mathbf{G}_j \mathbf{Z}_j' + \mathbf{R} = \mathbf{ZGZ}' + \mathbf{R},$$

où $\mathbf{Z} = [\mathbf{Z}_1 | \dots | \mathbf{Z}_s]$. On présume généralement que les matrices \mathbf{G}_j, \mathbf{R} dépendent d'un vecteur \mathbf{k} des composantes de variance ψ , ce qui nous permet d'écrire : $V(\psi) = \mathbf{ZG}(\psi)\mathbf{Z}' + \mathbf{R}(\psi)$.

Soulignons qu'au niveau des observations individuelles, le modèle (4) peut être révisé comme suit : $y_{dh} = \mathbf{x}_{dh}'\beta + \mathbf{z}_{dh1}'\mathbf{v}_1 + \dots + \mathbf{z}_{dhs}'\mathbf{v}_s + e_{dh}$.

Nous tenons compte de diverses spécifications pour les modèles linéaires mixtes, qui peuvent aussi être considérés comme des cas spéciaux du modèle général (4). Par souci de simplicité, nous avons adopté la notation au niveau des unités pour décrire les modèles à l'étude. Le premier

$$(5) \quad MM1 : y_{dh} = \mathbf{x}_{dh}'\beta + v_d + \alpha_i + e_{dhi},$$

peut être obtenu à partir de la formule (4) en écrivant $s = 2$, $q_1 = m$, $q_2 = 1$, $\mathbf{G}_1 = \sigma_v^2 \mathbf{I}_m$, $\mathbf{G}_2 = \sigma_\alpha^2 \mathbf{I}_I$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Ce modèle tient compte des effets régionaux et temporels indépendants les uns des autres, ce qui signifie qu'on présume que les effets régionaux n'évoluent pas au fil du temps. Cette structure d'effets aléatoires correspond à l'hypothèse d'une covariance constante entre les unités d'une même région, observée à deux moments donnés. Le deuxième modèle :

$$(6) \quad MM2 : y_{dhi} = \mathbf{x}_{dhi}'\beta + \delta_i + e_{dhi},$$

correspond au cas particulier où $s = 1$, $q_1 = m$, $\mathbf{G}_1 = \sigma_\delta^2 \mathbf{I}_{mq}$, $\mathbf{R} = \sigma_e^2 \mathbf{I}_n$. Nous tenons compte des effets de l'interaction entre les régions et les périodes, c'est-à-dire que certains effets régionaux ne sont pas constants au fil du temps.

$$\phi^{DT} = \frac{EQM^D(\overline{y}^{DT, RSYN}) + EQM^D(\overline{y}^{DT, DIR})}{EQM^D(\overline{y}^{DT, RSYN})}$$
et EQM^D signifie que l'erreur quadratique moyenne (EQM) est évaluée par rapport à la distribution de la randomisation. La sélection de ϕ^{DT} donne lieu aux estimateurs composites $\overline{y}^{DT, COMP}$ qui sont à peu près optimaux à l'égard de l' EQM^D (voir Rao 2003, section 4.3). En pratique, les quantités de la formule de ϕ^{DT} sont inconnues et peuvent être estimées à partir des données. On peut obtenir des estimateurs sans biais et convergents pour l' $EQM^D(\overline{y}^{DT, DIR}) = V^D(\overline{y}^{DT, DIR})$ en utilisant des formules standard. Un estimateur approximatif sans biais fondé sur le plan de sondage de l' $EQM^D(\overline{y}^{DT, RSYN})$ peut être obtenu au moyen des formules proposées dans le livre de Rao (2003, section 4.2.4). En particulier, nous calculons l'approximation comme suit :

$$eqm^D(\overline{y}^{DT, RSYN}) \approx (\overline{y}^{DT, RSYN})^2 - V^D(\overline{y}^{DT, DIR}),$$

où eqm^D et V^D représentent les estimateurs des valeurs EQM^D et V^D correspondantes. Plus précisément, V^D est l'estimateur sans biais selon le plan de sondage ordinaire de V^D lorsque nous calculons sa moyenne par rapport à d , comme d'habitude, pour obtenir un estimateur plus stable. En fait, l'un des problèmes associés à mse_D est qu'il peut même être négatif. De plus, afin d'améliorer la fiabilité de l'estimateur, on peut avoir recours à un estimateur direct modifié empruntant de l'information d'une région à l'autre pour l'estimation du coefficient de régression. Lorsque de l'information auxiliaire est disponible, l'estimateur de régression généralisée (GREG),

$$(3) \quad \overline{y}^{DT, GREG} = \mathbf{X}'^{DT}\hat{\beta}^w + \frac{\sum_{j \in ST} w_j e_j}{\sum_{j \in ST} w_j},$$

corrige approximativement le biais de l'estimateur synthétique au moyen du terme $(\sum_{j \in ST} w_j)^{-1} \sum_{j \in ST} w_j e_j$, en fonction des résidus de régression e_j .

2.2 Les estimateurs fondés sur le modèle

Les estimateurs fondés sur le modèle que nous avons envisagés sont axés sur la spécification de modèles explicites pour les données de l'échantillon, qui sont stimulées par un processus hypothétique de génération des données. Par conséquent, le problème associé à l'estimation de \overline{y}^{DT} se résume à un problème de prédiction. Par ailleurs, les erreurs quadratiques moyennes et les autres propriétés statistiques des estimateurs sont habituellement évaluées par rapport au processus de génération des données. Nous nous sommes concentrés sur le modèle « au niveau des unités » en fonction de modèles rattachant y_{dhi} à un vecteur de covariables \mathbf{x}_{dhi} . L'utilisation de modèles explicites comporte

À la section 2, nous présentons une organisation générale de l'estimation pour petits domaines au moyen des données de l'enquête par panel et examinons brièvement les méthodes d'estimations pour petits domaines fondées sur le plan de sondage et sur le modèle. Dans cette section, nous élaborons les meilleurs prédicteurs linéaires sans biais empiriques (MPLSBE) et leurs estimateurs d'erreur quadratique moyenne (EQM) pour certains modèles transversaux et chronologiques au niveau des unités, grâce à la théorie disponible sur les MPLSBE en matière d'estimation pour petits domaines (pour plus de détails, voir Rao 2003, et Jiang et Lahiri 2006a). Soulignons que les modèles transversaux et chronologiques ont été envisagés dans la documentation sur les petits domaines, mais uniquement dans le contexte de la modélisation au niveau de la région (voir Rao et Yu 1994; Ghosh, Nangia et Kim 1996; Datta, Lahiri, Maity et Lu 1999; Datta, Lahiri et Maity 2002; Pfeffermann 2002, entre autres).

À la section 3, nous passons brièvement en revue le PCM et décrivons comment nous servons des données de cette enquête pour effectuer une étude par simulation Monte Carlo, en vue de comparer différents estimateurs pour petits domaines et leurs estimateurs d'EQM. À la section 4, nous faisons état des résultats de la simulation Monte Carlo. Nous mentionnons que la simulation avait pour objet d'évaluer les propriétés fondées sur le plan de sondage de tous les estimateurs, même s'ils sont dérivés en tant que prédicteurs fondés sur le modèle. Nous constatons que les MPLSBE donnent de très bons résultats comparativement aux estimateurs fondés sur le plan de sondage, même si notre pseudo-population comporte des signes de non-normalité. Toutefois, l'effet de la non-normalité de la pseudo-population semble avoir un effet sur à l'efficacité des estimateurs de l'EQM. Dans notre simulation, les estimateurs de l'EQM par la série de Taylor (voir Prasad et Rao 1990; Datta et Lahiri 1999, entre autres) et par la méthode bootstrap paramétrique (voir Butar et Lahiri 2003) se révèlent plus vulnérables à la non-normalité que la méthode jackknife de Jiang, Lahiri et Van (2002). Nous concluons le document avec quelques observations finales.

2. Les méthodes d'estimation sur petits domaines considérées

Pour décrire les données de l'échantillon, y_{dnt} indique la valeur d'une variable de l'étude pour l'unité i appartenant à la région d pendant la période t ($d = 1, \dots, m$; $t = 1, \dots, T$; $i = 1, \dots, n_d$). De plus, \mathbf{x}_{dnt} correspond aux valeurs du vecteur des covariables associées à chaque y_{dnt} (et dont le premier élément est égal à 1), et $\mathbf{X}_s = \{\mathbf{x}_{dnt}^s\}$ représente la matrice $n \times p$ des valeurs des covariables pour l'échantillon au complet ($n = \sum_{d=1}^m n_d$). Supposons que nous voulons

prédire les moyennes régionales de la variable cible pendant la période finale T : \bar{Y}^{TP} , ($d = 1, \dots, m$). Supposons également que les vecteurs des valeurs de la population moyenne des covariables sont connus pour la période T ; nous désignons ces vecteurs comme suit: \mathbf{X}'^{TP} .

2.1 Estimateurs fondés sur le plan de sondage

Une première solution au problème de l'estimation pour petits domaines consiste à utiliser des estimateurs directs, c'est-à-dire des estimateurs utilisant seulement les valeurs y obtenues de la région (et de la période) auxquelles se rapporte le paramètre. L'estimateur direct le plus simple de la moyenne de population est la moyenne pondérée. Nous désignons cet estimateur direct comme suit: $\bar{y}^{PT, DIR}$ ($d = 1, \dots, m$), et nous nous en servons comme point de repère dans les sections qui suivent.

On peut généralement définir les estimateurs synthétiques comme des estimateurs sans biais pour une plus grande région comportant des erreurs-types acceptables. Ces estimateurs servent à calculer les estimations pour petits domaines, en supposant que les petits domaines ont les mêmes caractéristiques que les plus grands. En outre, lorsqu'il existe de l'information au sujet des variables auxiliaires, on peut obtenir un estimateur synthétique en particulier (l'estimateur de régression) en appliquant un modèle de régression à toutes les données de l'échantillon. Soulignons que l'estimateur synthétique se rapporte à une région par rapport aux variables auxiliaires et non par rapport à la variable à l'étude.

Par exemple, si nous considérons uniquement les observations du dernier cycle ($t = T$), le modèle de régression simple serait formulé comme suit:

$$y^{PT} = \mathbf{x}'^{PT} \beta + e^{PT} \quad E(e^{PT}) = 0, \quad E(e^{PT})^2 = \pi^2.$$

Pour tenir compte de la complexité du plan d'échantillonnage, on peut obtenir l'estimation pondérée par les moindres carrés β^w de β , ce qui signifie que l'estimateur de régression synthétique sera calculé comme suit:

$$(1) \quad \bar{y}^{PT, RSYN} = \mathbf{X}'^{PT} \beta^w, \quad d = 1, \dots, m.$$

En général, les estimateurs synthétiques ont des variances très faibles, mais ils peuvent être entachés de forts biais lorsque le modèle qui s'applique à l'échantillon en entier n'est pas bien adapté aux données régionales. Les estimateurs composites sont des moyennes pondérées des estimateurs direct et synthétique. Soit la formule suivante de l'estimateur composite:

$$(2) \quad \bar{y}^{PT, COMP} = \phi^{PT} \bar{y}^{PT, DIR} + (1 - \phi^{PT}) \bar{y}^{PT, RSYN},$$

Estimation pour petits domaines du revenu moyen des ménages en fonction des modèles au niveau des unités pour les données d'enquêtes par panel

Enrico Fabrizi, Maria Rosaria Ferrante et Silvia Pacci

Résumé

Le panel communautaire de ménages (PCM) est une enquête par panel qui porte sur un large éventail de sujets concernant les conditions socio-économiques et les conditions de vie. Plus précisément, cette enquête permet de calculer le revenu équivalent disponible des ménages, qui constitue une variable clé de l'étude de l'inégalité économique et de la pauvreté. Pour obtenir des estimations fiables de la moyenne de cette variable pour des régions données de pays, il faut avoir recours aux méthodes d'estimation pour petits domaines. Dans le présent document, nous nous attardons sur les prédicteurs linéaires empiriques du revenu équivalent moyen en fonction de l'emprunt d'information des « modèles au niveau des unités », d'une région à l'autre et d'une période à l'autre. En nous appuyant sur une étude par simulation basée sur les données du PCM, nous comparons les estimateurs suggérés avec les estimateurs transversaux, fondés sur les modèles et fondés sur le plan de sondage. Dans le cas de ces prédicteurs empiriques, nous comparons également trois différents types d'estimateurs de l'EQM. Les résultats indiquent que les estimateurs qui sont rattachés aux modèles qui tiennent compte de l'autocorrélation des unités entraînent d'importants gains d'efficacité, même en l'absence de covariables dont on connaît la moyenne de population.

Mots clés : Panel communautaire de ménages; revenu équivalent moyen; modèles linéaires mixtes; prédicteur linéaire sans biais empirique; estimation de l'EQM.

1. Introduction

Depuis quelques années, le milieu universitaire s'intéresse de plus en plus à l'analyse des disparités économiques régionales, qui entraînent sérieusement la croissance économique nationale, et donc, la cohésion sociale. Ce phénomène est particulièrement présent au sein de l'Union européenne, où les disparités régionales sont un des traits distinctifs de l'économie. Ce regain d'intérêt envers les économies locales a attiré la demande en matière d'information statistique régionale et stimulé les recherches sur la répartition du revenu, la pauvreté et l'exclusion sociale au niveau international.

Dans les années 90, Eurostat (l'organisme statistique de l'UE) a lancé le Panel communautaire de ménages (PCM), une enquête annuelle par panel menée auprès des ménages européens au moyen de méthodes normalisées dans tous les pays membres de l'UE (Betti et Verma 2002; Eurostat 2002). Le PCM a pris fin en 2001, après huit cycles. À l'heure actuelle, on est en train de le remplacer par l'Enquête européenne sur le revenu et les conditions de vie (EU-SILC), qui ressemble au PCM à bien des égards, mais dont on n'a pas encore publié les données. Le PCM portait sur un large éventail de sujets et permettait entre autres de calculer le revenu équivalent disponible des ménages, qui constitue une variable clé de l'étude de l'équité économique et de la pauvreté.

Le PCM avait pour objet de fournir des estimations fiables pour les grandes régions des pays de la NUTS1 (NUTS renvoie à la « Nomenclature des unités territoriales statistiques », qui est définie en fonction de certains principes décrits sur le site Web d'Eurostat au http://europa.eu.int/comm/Eurostat/ramon/Noix/home_regions_en.htm). Malheureusement, la NUTS1 renvoie à des régions (cinq groupes de régions administratives dans le cas de l'Italie) qui sont trop vastes pour permettre la mesure efficace de la disparité du revenu des régions ou pour fournir des renseignements utiles aux fins de la gouvernance régionale. Par conséquent, pour obtenir des estimations à un niveau géographique plus précis, on a employé une méthode d'estimation pour petits domaines. Le problème réside dans le choix d'une méthode appropriée et efficace.

Dans le présent document, afin d'amalgamer les renseignements provenant des enquêtes précédentes, les variables auxiliaires connexes et les petites régions, nous envisageons plusieurs extensions possibles du modèle de régression à erreurs emboîtées au niveau des unités, qui est bien connu (voir Bates, Harter et Fuller 1988), pour estimer le revenu équivalent moyen des ménages. En nous appuyant sur les données du PCM, nous illustrons comment les modèles proposés pourraient s'avérer utiles et améliorer l'efficacité des estimations pour petits domaines en misant sur la corrélation du revenu des ménages individuels au fil du temps.

1. Enrico Fabrizi, DMSIA, University of Bergamo, via del Caniana 2, 24127, Bergamo, Italy; Courtiel : enrico.fabrizi@unibg.it; M.R. Ferrante, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy; Courtiel : ferrante@stat.unibo.it; S. Pacci, Department of Statistics, University of Bologna, via Belle Arti 41, 40126, Bologna, Italy; Courtiel : pacci@stat.unibo.it.

- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Methodological*, Séries B, 36, 192-236.
- Carlin, B.P., et Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall Ltd.
- Clayton, D., et Kalbfleisch, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.
- Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc. Troisième édition.
- Das, K., Jiang, J., et Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.
- Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. Dans *New developments in survey sampling*, (Eds., N.L. Johnson et H.J. Smith). New York : Wiley-Interscience, 629-651.
- Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, A., Carlin, J., Stern, H. et Rubin, D. (1995). *Bayesian Data Analysis*. Chapman & Hall, London, U.K.
- Ghosh, M., Natarajan, K., Stroud, T. et Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Gillies, W.R., Richardson, S. et Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Gillies, W.R., et Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41, 337-348.
- He, C.Z., et Sun, D. (1998). Hierarchical Bayes estimation of hunting success rates. *Environmental and Ecological Statistics*, 5, 223-236.
- He, C.Z., et Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.
- Heidelberger, P., et Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.
- Jiang, J., et Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15, 1-96.
- Jiang, J., Lahiri, P. et Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with *M*-estimation. *The Annals of Statistics*, 30, 1782-1810.
- Kim, H., Sun, D. et Tsurakawa, R.K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, 96, 1506-1521.
- Kimmel, R.O. (2001). Regulating spring wild turkey hunting based on population and hunting quality. Dans *Proceeding of the National Wildlife Turkey Symposium*, 8, 243-250.
- Lohr, S.L. (1999). *Sampling: Design and analysis*. Duxbury Press.
- Malec, D., Sedransk, J., Mortaniry, C.L. et LeClerc, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.
- Olesen, J.J., et He, C.Z. (2004). Space-time modeling for the Missouri Turkey Hunting Survey. *Environmental and Ecological Statistics*, 11, 85-101.
- Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.
- Rao, J.N.K. (2003). *Small Area Estimation*. Wiley-Interscience.
- Rao, J.N.K. (2005). Inferential issues in small area estimation. Some new developments. *Statistics in Transition*, 7, 513-526.
- Robert, C.P., et Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.
- Smith, B.J. (2005). Bayesian Output Analysis program (BOA), version 1.1.5. <http://www.publhc-health.uiowa.edu/boa>.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. et Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (Pkg: 583-639). *Journal of the Royal Statistical Society, Methodological*, Séries B, 64, 583-616.
- Sun, D., Speckman, P.L. et Tsurakawa, R.K. (2000). Random effects in generalized linear mixed models (GLMMs). Dans *Generalized Linear Models: A Bayesian Perspective*, (Eds., D.K. Dey, S.K. Ghosh et B.K. Mallick). New York : Marcel Dekker, 23-39.
- Vangilder, L.D., Sherriff, S.L. et Olesen, G.S. (1990). Characteristics, attitudes, and preferences of Missouri's spring turkey hunters. Dans *Proceeding of the National Wildlife Turkey Symposium*, 6, 167-176.
- Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Statistical Methodology*, Séries B, 62, 159-180.
- Woodard, R., He, C.Z. et Sun, D. (2003). Bayesian estimation of hunting success rate and harvest for spatially correlated post-stratified data. *Biometrical Journal*, 45, 985-1005.
- You, Y., et Rao, J. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 111, 197-208.

Dans la fonction de perte par erreur quadratique, le meilleur « prédicteur » est alors la moyenne prévisionnelle donnée par

$$E\{y_{ik}^*|\mathbf{d}\}=\int\left\{\int_{y_{ik}^*}y_{ik}^*g_3(y_{ik}^*|\xi,\mathbf{d})dy_{ik}^*\right\}[\xi|\mathbf{d}]d\xi$$

De même, le meilleur « prédicteur » de $h(y_{ip}^*,\dots,y_{ik}^*)$ étant donné \mathbf{d} est

$$E\{h(y_{ip}^*,\dots,y_{ik}^*)|\mathbf{d}\}=\int E\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi,\mathbf{d}\}[\xi|\mathbf{d}]d\xi,\quad (14)$$

où

$$E\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi,\mathbf{d}\} =$$

$$\int_{(y_{ip}^*,\dots,y_{ik}^*)}h(y_{ip}^*,\dots,y_{ik}^*)\prod_{k=1}^K[y_{ik}^*|\xi,\mathbf{d}]dy_{ip}^*\dots dy_{ik}^*.\quad (15)$$

Pour la famille de distributions $g_3(y_{ik}^*|\xi,\mathbf{d})$, le côté droit de (15) est souvent une expression de forme fermée.

On peut aisément calculer les valeurs de prévision

bayésiennes de (14) en se fondant sur un échantillon aléatoire

tiré de la codistribution a posteriori de ξ . Ainsi, soit $\xi_{(l)}, l=1,\dots,L$, le produit de la simulation MCMC. On peut alors approcher la moyenne prévisionnelle a posteriori

(14) par

$$\hat{E}\{h(y_{ip}^*,\dots,y_{ik}^*)|\mathbf{d}\}=\frac{1}{L}\sum_{l=1}^LE\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi_{(l)},\mathbf{d}\}.$$

La variance prévisionnelle a posteriori de $h(y_{ip}^*,\dots,y_{ik}^*)$ étant donné \mathbf{d} peut ainsi s'exprimer :

$$V\{h(y_{ip}^*,\dots,y_{ik}^*)|\mathbf{d}\}=E[V\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi,\mathbf{d}\}|\mathbf{d}]$$

$$+V[E\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi,\mathbf{d}\}|\mathbf{d}].$$

On peut aussi approcher la variance prévisionnelle par

$$V\{h(y_{ip}^*,\dots,y_{ik}^*)|\mathbf{d}\}=\frac{1}{L}\sum_{l=1}^LV\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi_{(l)},\mathbf{d}\}$$

$$+\frac{1}{L}\sum_{l=1}^TE\{h(y_{ip}^*,\dots,y_{ik}^*)|\xi_{(l)},\mathbf{d}\}-E\{h(y_{ip}^*,\dots,y_{ik}^*)|\mathbf{d}\}^2.$$

4. Observations

Dans cet article, nous avons conçu un MMLG à deux variables pour deux paramètres canoniques inconnus. Il le fallait pour obtenir des estimations lorsque la taille d'échantillon est aléatoire et qu'on a besoin d'estimations de N_{ik}^* en sus d'estimations de η_{ik} . Nous avons élaboré notre modèle à l'aide de deux MMLG simultanés dans une structure hiérarchisée de Bayes. Le modèle que nous proposons a pour avantage d'être applicable à une grande

diversité de problèmes. L'introduction d'une taille d'échantillon aléatoire et l'estimation des tailles de population représentées des techniques utiles dans bien des applications.

Bien sûr, nous pensons qu'il y a un rapport inverse entre les taux de succès de la chasse et le nombre d'expéditions dans un comté. Pour modéliser le phénomène, nous pouvons songer à un modèle ARC double de représentation des relations entre η_{ik} et N_{ik}^* comme le modèle de Kim, Sun et Tsutakawa (2001). Un modèle ARC à plus de deux variables est peut-être un autre moyen à notre disposition.

Pour chacun des modèles spatiaux, nous avons posé une corrélation commune ρ_k à l'échelle de l'État. On pourrait juger plus raisonnable d'introduire des termes supplémentaires de corrélation dans différentes régions de productivité délimitées par les services de conservation du Missouri. Ce serait là un apport intéressant mais complexe au modèle. Il faut également dire que la structure spatiale que nous avons utilisée est semblable à celle de He et Sun (2000) et d'Oleson et He (2004), études où cette modélisation spatiale a donné de bons résultats.

Il serait peut-être bon de tenir compte de la distance entre le lieu de résidence et le lieu de chasse des chasseurs pour le lieu de résidence et s'éloignent pas trop de leur lieu de résidence et une telle information pourrait être intégrée au cadre hiérarchisé.

À noter que l'estimation de la récolte est supérieure au dénombrement effectué aux postes de contrôle. Une explication partielle en est que les chasseurs qui réussissent mieux ont tendance à répondre au questionnaire postal. Nous faisons une recherche pour corriger le biais de non-réponse.

Remerciements

Cette étude a été partiellement soutenue par le projet W-13-R du programme d'aide au rétablissement de la faune du gouvernement fédéral américain. Pour ses travaux, Sun a reçu la subvention SES-0351523 de la National Science Foundation et la subvention R01-CA100760 de la NIH. Les auteurs aimeraient remercier de leurs précieuses observations le D^rLarry Vangilder et M. Jeff Berthnger des services de conservation de l'État du Missouri. Ils remercient enfin vivement le rédacteur en chef, un rédacteur adjoint et deux examinateurs pour leurs commentaires constructifs sur les versions a priori de la présente étude.

Bibliographie

Agresti, A. (2002). *Categorical data analysts*. New York : John Wiley & Sons, Inc.

(iii) $\theta^{(w)}$ sont mutuellement indépendants et pré-sentent les distributions a posteriori conditionnelles

(iv) $(\cdot | \mathbf{n})^{\mathcal{Y}}$ sont mutuellement indépendants et pré-sentent les distributions a posteriori conditionnelles

$$\left(\begin{array}{c} \left(\mathbf{X}_i^v \mathbf{X}_i^v \frac{\gamma^v \mathbf{g}}{\mathbf{I}} + \mathbf{I} \frac{\gamma^v \mathbf{1}}{\mathbf{I}} \right) \\ \mathbf{I}^{-1} \left(\mathbf{I} \frac{\gamma^v \mathbf{1}}{\mathbf{I}} + (\gamma^v \mathbf{n} - \gamma^v \mathbf{A})_i \mathbf{X}_i^v \frac{\gamma^{(v)} \mathbf{g}}{\mathbf{I}} \right) \\ \left(\mathbf{X}_{iJ}^v \mathbf{X}_{iJ}^v \frac{\gamma^v \mathbf{g}}{\mathbf{I}} + \mathbf{I} \frac{\gamma^v \mathbf{1}}{\mathbf{I}} \right) \end{array} \right) \mathbf{I}^{-1} N$$

$$(V) \quad (G_{(\varepsilon)}^{\mathcal{K}} | \cdot) \text{ ont les distributions a posteriori gamma inverses}$$

$$I_N \begin{pmatrix} \frac{1}{I} \mathbf{S}_a^a \mathbf{S}_a^a \frac{1}{I} \mathbf{Q}_{(a)}^{(a)} + \frac{1}{I} \mathbf{B}^{ak} \frac{1}{I} \mathbf{Q}_{(n)}^{(n)} \\ \frac{1}{I} \mathbf{S}_a^a \mathbf{S}_a^a \frac{1}{I} \mathbf{Q}_{(a)}^{(a)} - \mathbf{V}^{ak} \mathbf{X}^{ak} \theta^{ak} \\ \frac{1}{I} \mathbf{S}_a^a \mathbf{S}_a^a \frac{1}{I} \mathbf{Q}_{(a)}^{(a)} + \frac{1}{I} \mathbf{B}^{ak} \frac{1}{I} \mathbf{Q}_{(n)}^{(n)} \end{pmatrix}$$

$$(VI) \quad (\delta_{(n)}^{ak} | \cdot) \sim \text{Inverse Gamma}(\alpha_{(n)}^{ak} + I/2, \beta_{(n)}^{ak} + I/2 \mathbf{u}^{ak}) \quad \mathbf{B}^{ak} \mathbf{u}^{ak}.$$

$$\text{(vii)} \quad |p^{ak}| \propto |B^{ak}|^{1/2} \exp \left\{ -\frac{u'_{ak} B^{ak} n^{ak}}{28^{(n)}_{ak}} \right\}.$$

(viii) Si ϕ_1 a la densité a priori $p(\phi_1)$,

$$\left\{ \begin{aligned} &A_1(\phi_1)[v_{1ik}^{-1}h_1^{1ik}] + C_1(v_{1ik}, u_{1ik}) \\ &- B_1(h_1^{1ik}, v_{1ik}) \end{aligned} \right\} \exp(\phi_1) \propto [\cdot] \phi_1$$

(ix) Si ϕ_2 a la densité a priori $p(\phi_2)$,

$$\left\{ \begin{array}{l} A_2(\phi^{(2)}_1[n]y^{(2)}_{1k}) \\ -B_2(\phi^{(2)}_1v^{(2)}_{1k})) \\ C_2(\phi^{(2)}_1, \phi^{(2)}_2) \end{array} \right\} \text{exp} \{ [\cdot | \phi] \phi^{(2)}_2 \} \propto [\cdot | \phi] \phi^{(2)}_2$$

Lorsqu'on examine les parties (i), (ii) et (vii) du « Fait I », les densités conditionnelles de η_{ik} , ω_{ik} et p_{ak} sont souvent logconcaves.

sont souvent logconcaves.

3.3 Estimations quantitatives des domaines d'étude

Il est souvent intéressant d'estimer certaines quantités relatives aux domaines d'étude. Pour estimer, par exemple, une quantité du domaine i , prenons

$$(\omega^K, u^K; \dots; \omega^1, u^1) f = \psi$$

On peut aisément calculer des estimations bayésiennes de ψ_l en se fondant sur un échantillon aléatoire de la codistribution a posteriori. Soit $(\pi_l^{(k)})_{k=1}^K$, $l = 1, \dots, L$ et $k = 1, \dots, K$ le produit de simulation MCMC et la définition

$$\cdot \binom{K!}{(l)} \omega, \binom{K!}{(l)} \omega, \dots, \binom{l!}{(l)} \omega, \binom{l!}{(l)} \omega \Big) f = \binom{l!}{(l)} \hbar$$

Étant donné $y = \{y_{ik}, i = 1, \dots, I, k = 1, \dots, K\}$, avec la moyenne et la variance a posteriori de ψ , on peut approcher les formes générales de (4) et (5) par

$$({}^I\mathfrak{h} \mid \sum_{\tau=1}^{I-1} \frac{\tau}{I}) = (\mathfrak{A} \mid {}^I\mathfrak{h}) \underline{\mathfrak{F}}$$

et

$$z\{(\mathcal{A} \mid \mathcal{H})_{\mathcal{F}}\} \frac{1-T}{T} - z\{(\mathcal{H})\} \sum_{\mathcal{I}} \frac{1-T}{1} = (\mathcal{A} \mid \mathcal{H})_{\mathcal{A}}$$

respectively.

3.4 Prévisions quantitatives des domaines d'étude

Soit N_{ik} la taille de population et Y_{ik}^* la valeur d'intérêt pour les chasseurs non échantillonnés du domaine d'étude ?

quantités $Y_i^* = \sum_{k=1}^K Y_{ik}^*$, la valeur totale de réponse dans le domaine d'étude i . Pour une valeur donnée de n_{ik} , Y_{ik}^* devrait être de la même famille que Y_{ik} , de sorte que

$$\mathcal{S}_3(V^{\frac{1}{2}k} | n^{\frac{1}{2}k}, n^{\frac{1}{2}k}, N^{\frac{1}{2}k}, \phi_1) = \exp \left[\begin{aligned} & A_1(\phi_1) \{ V^{\frac{1}{2}k} \} \{ n^{\frac{1}{2}k} \} \\ & - B_1(n^{\frac{1}{2}k}, N^{\frac{1}{2}k} - n^{\frac{1}{2}k}) \\ & + C_1(V^{\frac{1}{2}k}, N^{\frac{1}{2}k} - n^{\frac{1}{2}k}, \phi_1) \end{aligned} \right]$$

C'est la simplement (9), n_k étant remplacé par $N_k - n_k$. Pour simplifier la notation, soit ξ les paramètres du modèle et \mathbf{d} les données. (Dans ce cas, $\mathbf{d} = \{(Y_k^i, n_k^i) : i = 1, \dots, K, k = 1, \dots, K\}$ et ξ pourrait comprendre les paramètres de la distribution a priori $\pi(\xi)$ (propre ou impropre), nous obtenons la densité a postérieur $[\xi | \mathbf{d}] \propto f(\mathbf{d} | \xi) \pi(\xi)$. Posons qu'une nouvelle observation Y_k^* suit $\mathcal{G}_3(Y_k^* | \mathbf{d}, \xi)$, qui pourrait dépendre de \mathbf{d} . La densité prévisionnelle de Y_k^* étant donnée \mathbf{d} se formule ainsi :

$$\mathbb{E}_p[p | \mathfrak{L}](p, \mathfrak{L} | \mathcal{Y}_*^k \mathcal{V}_*^{\mathfrak{L}} \mathfrak{L}) = [p | \mathcal{Y}_*^k \mathcal{V}_*^{\mathfrak{L}}]$$

3. Formules générales pour un MMLG

à deux variables

3.1 Modèle général

Nous ne tenons pas compte de la composante temporelle dans le modèle général par souci de simplicité. Soit n_k et Y_k les tailles d'échantillon et les variables d'intérêt pour le domaine d'étude i et la strate nominale k respectivement. Nous posons que $\{Y_k, n_k\}$ étant donné les paramètres inconnus $\{\eta_k, \omega_k, \phi_k, \psi_k\}$, avec $i = 1, \dots, I$, et $k = 1, \dots, K$ sont indépendants. Nous posons aussi que la fonction de densité conditionnelle de Y_k étant donné n_k appartient à la famille suivante de fonctions de densité de probabilité :

$$g_1(Y_k | \eta_k, n_k, \phi) = \exp[A_1(\phi_1)\{Y_k, n_k - B_1(\eta_k, n_k) + C_1(Y_k, n_k, \phi_1)\}], \quad (9)$$

où η_k est un paramètre inconnu. On forme souvent l'hypothèse que le paramètre d'échelle ϕ_1 est connu. La fonction de densité de probabilité de n_k appartient à la famille

$$g_2(n_k | \omega_k, \psi) =$$

$$\exp[A_2(\psi_2)\{n_k, \omega_k - B_2(\omega_k) + C_2(n_k, \psi_2)\}], \quad (10)$$

où ω_k est un paramètre inconnu égal à une fonction de taille de population N_k et où ψ_2 est souvent connu. La co-densité de Y_k et n_k (MMLG à deux variables) est alors

$$p(Y_k, n_k | \eta_k, \omega_k, \phi, \psi) = g_1(Y_k | \eta_k, n_k, \phi) g_2(n_k | \omega_k, \psi). \quad (11)$$

La famille de distributions (10) est souvent appelée « modèle linéaire généralisé » et comporte des distributions binomiales, de Poisson, normales et gamma (voir, par exemple, Sun et coll. 2000). La famille de distributions (9) est une généralisation d'un tel modèle linéaire généralisé par intégration d'un paramètre de plus. Quatre cas d'espèce de (9) sont les distributions binomiales, de Poisson, normales et gamma qui font toutes partie de la famille exponentielle. Le MMLG à deux variables est applicable lorsque les estimations offrent de l'intérêt au niveau des domaines et que la taille d'échantillon n_k est considérée à la fois comme relevant du hasard et de l'observation. Elle a aussi son utilité lorsque des estimations de N_k sont requises.

Un modèle mixte linéaire peut s'utiliser en distribution a priori de η_k pour rendre compte de sa variabilité. On pourrait toutefois s'intéresser tant à η_k qu'à ω_k ou à une fonction des deux, ω_k étant souvent fonction de la taille de population. Dans ce cas, il faudrait modéliser η_k et ω_k simultanément. Une catégorie générale de MMLG pour η_k pourrait être

$$h_1(\eta_k) = X_1' \theta_k + S_1' u_k + e_{1k} \quad (12)$$

Avec $a = 1, 2$, $X_a = \{x_{aik}\}$ et $S_a = \{s_{aik}\}$ sont des matrices connues de plan d'échantillonnage. Le vecteur θ_{ak} est le vecteur des effets fixes et u_{ak} , celui des effets variables; e_{aik} représente les effets résiduels indépendants et $e_{aik} \sim N(0, \delta_{aik}^2)$. On pose en outre que u_{ak} et e_{aik} indépendants l'un de l'autre.

3.2 Distributions a priori supplémentaires

Pour compléter le modèle hiérarchisé de Bayes, nous devons spécifier les distributions a priori pour $(\theta_{ak}, u_{ak}, \delta_{aik}^2)$, $a = 1, 2$, $k = 1, \dots, K$. La distribution a priori commune pour les effets fixes θ_{ak} est normale avec une variance importante ou une valeur a priori constante. Les effets aléatoires sont en corrélation spatiale dans bien des cas. La densité peut être de la forme ARC où la codensité est donnée par l'équation (3) avec $B_{ak} = I - \rho_{ak} C$. Mentionnons enfin qu'une distribution a priori commune des composantes de la variance δ_{aik}^2 est une distribution gamma inverse.

Dans l'évaluation de la distribution a posteriori, on peut employer des méthodes MCMC comme la formule d'échantillonnage de Gibbs pour tirer des échantillons de cette distribution. Nous présentons les distributions conditionnelles intégrales dans le cas général qui suit.

Fait 1 Soit (52) les distributions conditionnelles de Ω étant donné tous les autres paramètres et soit [52] la densité conditionnelle. Voici les densités a posteriori conditionnelles de η_k et ω_k :

$$(i) \quad [\eta_k | \cdot] \propto \exp\{A_1(\phi_1)[Y_k, \eta_k] - B_1(\eta_k, n_k) - 1/2\delta_{1k}^2 [h_1(\eta_k) - X_1' \theta_{1k} - S_1' u_{1k}]^2\} h_1^*(\eta_k) \quad \text{ou, d'une manière équivalente } \nu_{1k} = h_1(\eta_k) \text{ ayant la densité conditionnelle}$$

$$[v_{1k} | \cdot] \propto \exp\left\{A_1(\phi_1)[Y_k, h_1^{-1}(v_{1k})] - B_1(h_1^{-1}(v_{1k}), n_k) - 1/2\delta_{1k}^2 (v_{1k} - X_1' \theta_{1k} - S_1' u_{1k})^2\right\}$$

$$(iii) \quad [\omega_k | \cdot] \propto \exp\{A_2(\psi_2)[n_k, \omega_k] - B_2(\omega_k) - 1/2\delta_{2k}^2 [h_2(\omega_k) - X_2' \theta_{2k} - S_2' u_{2k}]^2\} h_2^*(\omega_k) \quad \text{ou, d'une manière équivalente, } \nu_{2k} = h_2(\omega_k) \text{ ayant la densité conditionnelle}$$

$$[v_{2k} | \cdot] \propto \exp\left\{A_2(\psi_2)[n_k, h_2^{-1}(v_{2k})] - B_2(h_2^{-1}(v_{2k}), n_k) - 1/2\delta_{2k}^2 (v_{2k} - X_2' \theta_{2k} - S_2' u_{2k})^2\right\}$$

Pour $a = 1, 2$, nous obtenons les distributions conditionnelles suivantes :

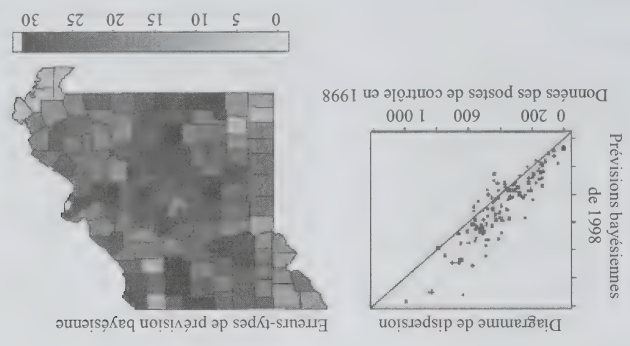
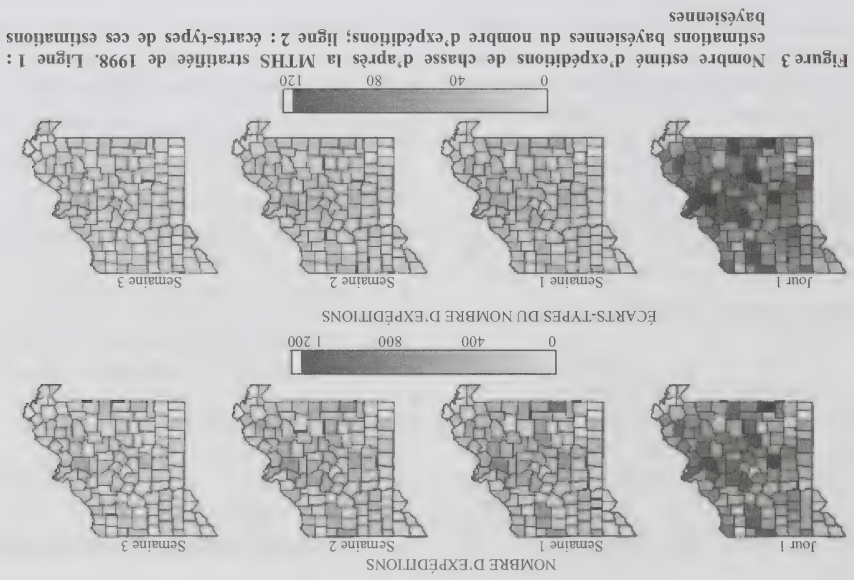
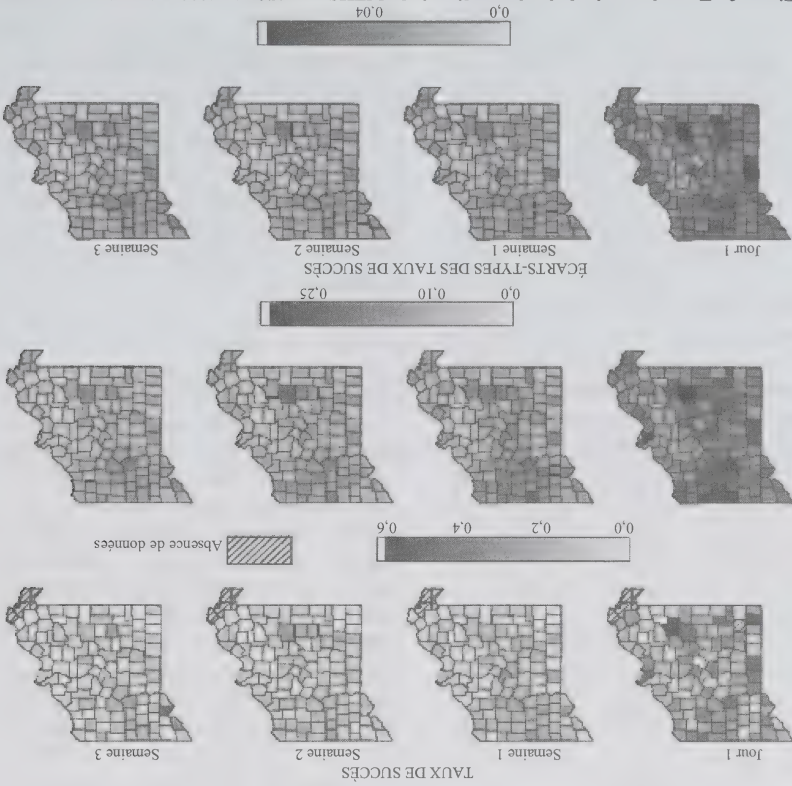


Figure 4 Données des postes de contrôle en 1998, prévisions bayésiennes, diagramme de dispersion de ces données selon les prévisions et erreurs-types de ces dernières

Figure 2 Taux de succès de la chasse d'après la MTHS stratifiée de 1998. Ligne 1 : estimations « nominales » des taux de réussite; ligne 2 : estimations bayésiennes de ces taux; ligne 3 : écarts-types des estimations bayésiennes des taux de succès



des postes de contrôle. À l'aide de la formule (8), nous La première carte de la figure 4 livre les données réelles saison de chasse.

chassaient la première journée que tout autre jour de la population N_{jk}^i . Les estimations de modélisation de N_{jk}^i figurent à la première ligne de la figure 3. Leurs erreurs-représentées pour les semaines 1, 2 et 3 sont les moyennes quotidiennes de ces semaines. Il est clair que plus de gens types sont présentes à la deuxième ligne. Les valeurs de la MTHS à l'aide des données de 1996 (He et Sun 2000; Olson et He 2004). Les estimations les plus hautes ne se

Nous produisons également des estimations de la taille de défensesurs du mouvement de conservation. une tendance temporelle, ainsi que le font observer les peu plus à l'est qu'ils ne l'étaient en 1996, ce que l'on doit à hauts taux d'estimation à l'aide des données de 1998 sont un présentement cependant pas dans les mêmes comtés. Les plus de la MTHS à l'aide des données de 1996 (He et Sun 2000; Olson et He 2004). Les estimations les plus hautes ne se

Les estimations tendent à décroître tout au long de la saison de chasse. Ajoutons que les estimations les plus élevées se rattachent à la partie nord de l'État du Missouri. C'est aussi ce qu'on a pu vérifier dans des analyses a priori de la MTHS à l'aide des données de 1996 (He et Sun 2000; Olson et He 2004). Les estimations les plus hautes ne se présentent cependant pas dans les mêmes comtés. Les plus hauts taux d'estimation à l'aide des données de 1998 sont un peu plus à l'est qu'ils ne l'étaient en 1996, ce que l'on doit à une tendance temporelle, ainsi que le font observer les défensesurs du mouvement de conservation. Nous produisons également des estimations de la taille de la population N_{jk}^i . Les estimations de modélisation de N_{jk}^i figurent à la première ligne de la figure 3. Leurs erreurs-représentées pour les semaines 1, 2 et 3 sont les moyennes quotidiennes de ces semaines. Il est clair que plus de gens types sont présentes à la deuxième ligne. Les valeurs de la MTHS à l'aide des données de 1996 (He et Sun 2000; Olson et He 2004). Les estimations les plus hautes ne se

IG(2,1), ce qui donne l'unité comme moyenne et l'infini comme variance. Nous donnons des valeurs a priori informatives à $\theta_{1jk} \sim N(-1.5, 16)$ et $\theta_{2jk} \sim N(4.25)$. Pour obtenir une distribution a priori informative, nous avons modélisé les deux étapes suivantes, parce qu'une distribution a priori informative qui serait entièrement arbitraire pourrait mener à des estimations a postériori peu sûres (Wasserman 2000). D'abord, nous avons pris des valeurs a priori non informatives (mais conjuguées) pour dégager des moyennes et des écarts-types a postériori pour chacun des paramètres considérés par des simulations MCMC. Nous avons ensuite posé que les moyennes a priori informatives étaient proches des moyennes a postériori de θ_{1jk} et θ_{2jk} et que les écarts-types a priori informatifs étaient environ décuplés des écarts-types a postériori une fois les estimations a postériori obtenues par distribution a priori non informative. Les estimations issues de ces deux manières de procéder étaient très convergentes, mais le modèle à distributions a priori informatives donnait des variantes moindres.

Nous avons ajusté un grand nombre de modèles simplifiés à des fins de comparaison et de contrôle de modélisation. Comme moyen de sélection de modèle, nous avons appliqué le critère d'information par l'écart quadratique proposé par Spiegelhalter, Best, Carlin et van der Linde (2002). C'est là une généralisation du critère d'information d'Akaike qui mesure les carrés d'écart pour une exécution de l'ajustement où la dimensionnalité du modèle entre en ligne de compte. Des valeurs plus basses du critère d'information par écart quadratique indiquent un meilleur ajustement. Notre modèle avec valeurs a priori informatives présente 13 910,5 comme valeur d'écart quadratique. Pour aucun autre modèle, la valeur d'écart n'était significativement réduite par rapport à ce modèle. Le modèle à valeurs a priori non informatives pour θ_{1jk} et θ_{2jk} se caractérise par une valeur d'écart de 14 700,4. Pour un modèle réduit à paramètres de corrélation commune $p_{11} = p_{12} = p_{13} = p_{14} = p_{15}$ et $p_{21} = p_{22} = p_{23} = p_{24} = p_{25}$, la valeur d'écart s'établit à 13 896,9.

Comme autre contrôle de modélisation, nous avons calculé les moyennes globales (dans tout l'État du Missouri) des estimations en prenant à la fois les simples estimations « natives » du plan d'échantillonnage et les estimations bayésiennes qui suivent au tableau 2. Au niveau de l'État, les tailles d'échantillon sont suffisamment grandes pour que les estimations basées sur le plan d'échantillonnage soient jugées fiables. Les estimations selon le plan correspondent de près aux estimations par modèle. Ainsi, l'estimateur de Bayes donne de bons résultats pour la propriété de cohérence avec les estimations selon le plan (voir You et Rao 2003, et Jiang et Lahiri 2006). Nous notons que les

dindons récoltés dans le comté i et la période j par les chasseurs non échantillonnés, soit y_{ijk}^* le nombre de dindons récoltés dans le comté i et la période j pour la strate k . Ainsi, le nombre de dindons récoltés dans le comté i et la période j par les chasseurs de la strate k est $h_{ijk} = y_{ijk}^* + y_{ijk}^*$. Dans ce cas, y_{ijk}^* est une valeur connue et nous avons seulement y_{ijk}^* à trouver. Nous pouvons voir y_{ijk}^* étant donné (n_{ijk}^*, p_{ijk}^*) comme une variable aléatoire binomiale ayant la forme de (1). Ainsi,

$$(y_{ijk}^* | n_{ijk}^*, p_{ijk}^*) \sim \text{Binomial}(n_{ijk}^*, p_{ijk}^*). \quad (6)$$

Soit $(p_{ijl}^{(0)}, n_{ijl}^{(0)})$, $l = 1, \dots, L$, le produit de l'application de la chaîne de Gibbs après une période initiale d'itérations. La moyenne prévisionnelle de y_{ijk}^* étant donné $\mathbf{p} = \{(y_{ijk}^*)^{jk} : i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K\}$ est alors

$$E(y_{ijk}^* | \mathbf{p}) = \frac{1}{L} \sum_{l=1}^L (n_{ijl}^{(0)} - n_{ijk}^*) p_{ijk}^*. \quad (7)$$

Enfin, nous écrivons

$$h_{ijk}^* = y_{ijk}^* + E(y_{ijk}^* | \mathbf{p}),$$

$$\hat{h}_i = \sum_{j=1}^J \sum_{k=1}^K \hat{h}_{ijk}^* \quad (8)$$

La variabilité de h_i est donnée par

$$V(h_i | \mathbf{p}) = \frac{1}{L} \sum_{l=1}^L \sum_{j=1}^J \sum_{k=1}^K (n_{ijl}^{(0)} - n_{ijk}^*) p_{ijk}^* (1 - p_{ijk}^*) + \frac{1}{L} \sum_{l=1}^L \sum_{j=1}^J \sum_{k=1}^K [V_{ijk}^* - n_{ijk}^* (p_{ijk}^*)^2] - h_{i(l)}^2 \Bigg\}^2.$$

La preuve est une application simple des espérances conditionnelles. À noter que h_{ijk}^* est défini comme variable aléatoire. Il n'est pas égal à $N_{ijk}^{(jk)} p_{ijk}^*$ pas plus que y_{ijk}^* ne l'est à $n_{ijk}^{(jk)} p_{ijk}^*$. On ne devrait donc pas se servir simplement de $N_{ijk}^{(jk)} p_{ijk}^*$ pour estimer h_{ijk}^* , bien que la valeur puisse être très proche si $N_{ijk}^{(jk)}$ est suffisamment élevée.

2.4 Ajustement du modèle

Nous avons exécuté 15 000 itérations pour la chaîne de Gibbs, dont 5 000 ont été écartées comme séquence initiale. Ainsi, les estimations a postériori sont fondées sur 10 000 échantillons autocorrélés de la distribution a postériori. Pour nous assurer la convergence de l'échantillonnage de Gibbs, nous avons recouru à la méthode de diagnostic de Heidelberger et Welch (1983), ainsi qu'à un contrôle graphique des parcours d'échantillon. Nous avons procédé à des diagnostics de convergence et à la détermination des valeurs sommariales a postériori au moyen du logiciel BOA (Smith 2005). Laissons les composantes de la variance $\delta_{(n)}^{(jk)}$ et $\delta_{(n)}^{(jk)}$ suivre une distribution a priori non informative avec

Pour cette distribution, la taille globale d'échantillon n_k est jugée aléatoire. Nous avons exposé à la section qui précède ce qui milite en faveur d'une telle hypothèse. Si n_k était fixe, la distribution multinomiale trait mieux comme modèle. Les valeurs de vraisemblance de ces deux traitements sont des plus convergentes et donnent des résultats comparables dans l'un et l'autre de ces cadres (voir Agresti 2002, pages 8-9).

Nous modélisons p_{ijk} par sa fonction logistiquie et N_{ijk} par une transformation logarithmique :

$$\eta_{ijk} = \log \left(\frac{1 - p_{ijk}}{p_{ijk}} \right), \quad \omega_{ijk} = \log(N_{ijk}).$$

Nous nous servons de modèles mixtes linéaires pour les distributions a priori à la fois dans η_{ijk} et ω_{ijk} en posant ce qui suit :

$$\eta_{ijk} = \theta_{ijk} + u_{ijk} + e_{ijk}, \\ \omega_{ijk} = \theta_{zjk} + u_{zjk} + e_{zjk}.$$

Pour $a = 1, 2$, $\theta_{ijk}^{(a)}$ désigne l'effet fixe du au j^{e} moment dans la strate k et $u_{ijk}^{(a)}$ représente l'effet aléatoire de comté; les erreurs aléatoires $e_{ijk}^{(a)}$ sont i.i.d. $N(0, \delta_{(a)}^2)$.

Pour compléter le modèle hiérarchisé de Bayes, nous devons spécifier les distributions a priori pour $\theta_{ijk}^{(a)} = (\theta_{ijk}^{(a)1}, \dots, \theta_{ijk}^{(a)K})$, $u_{ijk}^{(a)} = (u_{ijk}^{(a)1}, \dots, u_{ijk}^{(a)K})$ et $\delta_{(a)}^2$. He et Sun (2000) et Oleson et He (2004) démontrent l'existence d'un

$$f(\mathbf{u}^{(a)}) = \left\{ \exp \left[-\frac{1}{2\delta_{(a)}^2} (\mathbf{I} - \mathbf{p}^{(a)} \mathbf{C}^T \mathbf{C} \mathbf{I} - \mathbf{p}^{(a)} \mathbf{C}^T \mathbf{u}^{(a)} - \mathbf{p}^{(a)} \mathbf{C}^T \mathbf{u}^{(a)} \mathbf{C} \mathbf{I}) \right] \right\}.$$

Dans ce cas, $\delta_{(a)}^{1K}$ et $\delta_{(a)}^{2K}$ sont des composantes de la variance et $\mathbf{I} - \mathbf{p}^{(a)} \mathbf{C}^T$ est une matrice symétrique définie à valeurs non négatives (Besag 1974). \mathbf{I} est la matrice d'identité $\mathbf{I} \times \mathbf{I}$ et \mathbf{C} , une matrice d'adjacence dont la composante $\{c_{ij}\}$ est l'unité si les régions i et j sont en moyenneté avec $\{c_{ij}\} = 0$. Nous définissons également p_{ik} et p_{2k} comme paramètres de corrélation spatiale. Soit $\lambda_1 \leq \dots \leq \lambda_I$ les valeurs caractéristiques de la matrice \mathbf{C} . Ces matrices $\mathbf{I} - \mathbf{p}^{(a)} \mathbf{C}^T$, sont à valeurs définies positives si $\lambda_1^{-1} \leq p_{ak} \leq \lambda_I^{-1}$ (Clayton et Kaldor 1987). Pour les données du Missouri, $I = 114$ et les valeurs numériques de λ_1 et λ_{114} sont respectivement 2,8931 et 5,6938. Ainsi, il y a densité de $\mathbf{u}^{(a)}$ si p_{ak} se situe dans l'intervalle (-0,3457, 0,1756).

Pour les autres distributions a priori, nous posons des hypothèses. $\theta_{ijk}^{(a)}$ est normal avec une moyenne $\mu_{ijk}^{(a)}$ et une variance $\tau_{ijk}^{(a)}$. Soit p_{ak} en distribution unitforme sur

2.3 Estimation et prévision bayésiennes

Nous avons obtenu des estimations de $(p_{ijk}^{(a)}, N_{ijk}^{(a)})$. Nous désirons grouper les estimations de la strate pour pouvoir estimer $p_{ij}^{(a)}$ et $N_{ij}^{(a)}$. Nous voulons aussi prévoir le nombre inobservé de dindons récoltés dans le comté i , ce que désigne h_i .

Pour dégager des estimations de $(p_{ij}^{(a)}, N_{ij}^{(a)})$, soit $(p_{ij(i)}^{(a)}, N_{ij(i)}^{(a)})$, $i = 1, \dots, L$ le résultat de l'application de la chaîne d'échantillonnage de Gibbs après la période initiale d'itérations. Définissons

$$p_{ij(i)}^{(a)} = \frac{\sum_{k=1}^K p_{ijk}^{(a)} N_{ijk}^{(a)}}{\sum_{k=1}^K N_{ijk}^{(a)}}, \quad i = 1, \dots, L.$$

On peut alors approcher la moyenne et la variance de distribution a postérieure de $p_{ij}^{(a)}$ respectivement par

$$E(p_{ij}^{(a)}) = \frac{1}{L} \sum_{i=1}^L p_{ij(i)}^{(a)}, \quad V(p_{ij}^{(a)}) = \frac{1}{L} \sum_{i=1}^L \{p_{ij(i)}^{(a)}\}^2 - \left\{ \frac{1}{L} \sum_{i=1}^L p_{ij(i)}^{(a)} \right\}^2, \quad (4)$$

De même, définissons $N_{ij(i)}^{(a)} = \sum_{k=1}^K N_{ijk}^{(a)}$. On peut approcher la moyenne et la variance de distribution a postérieure de $N_{ij}^{(a)}$ à l'aide aussi du produit de la simulation MCMC.

Nous nous intéressons maintenant à la prévision bayésienne de la récolte inobservée de dindons par les distributions prévisionnelles a postérieures. Soit y_{ijk} le nombre de dindons récoltés dans le comté i , ce que désigne h_i .

ment proportionnelles à

l'intervalle $(\lambda_1^{-1}, \lambda_I^{-1})$. Enfin, une distribution a priori commune pour les composantes de la variance $\delta_{(a)}^{1K}$ et $\delta_{(a)}^{2K}$ est une distribution gamma inverse (Gelman, Carlin, Stern et Rubin 1995) dont les valeurs de densité sont respectivement proportionnelles à

$$\frac{1}{\beta_{(a)}^{\alpha_{(a)}}} \exp \left(-\frac{\delta_{(a)}^2}{\beta_{(a)}^{\alpha_{(a)}}} \right) \text{ et } \frac{1}{\beta_{(a)}^{\alpha_{(a)}}} \exp \left(-\frac{\delta_{(a)}^2}{\beta_{(a)}^{\alpha_{(a)}}} \right).$$

d'intérêt, parce que les chasseurs sont capables d'indiquer où ils ont chassé lorsqu'ils rendent compte du succès de leur

chasse. Les intéressés ont besoin d'un permis de chasse et, par conséquent, nous savons combien de chasseurs compte

le Missourï au total et à quelles strates ceux-ci appartiennent. Comme les données sont étalées sur les 114

comtés de cet État, on se retrouve avec une information fort

claire pour chaque comté. Plutôt que de regarder les

taux de succès de jour en jour, nous le faisons pour quatre

périodes que nous avons délimitées. Dans le calcul des taux

de succès par comté, nous nous reportons cependant au

nombre d'expéditions de chasse dans chaque comté. Nous

ignorons combien d'expéditions fera chaque chasseur ou

dans quel comté il les fera. Nous connaissons donc le

nombre de chasseurs dans chaque strate, mais le nombre

d'expéditions est inconnu. De plus, nous connaissons le

nombre de chasseurs échantillonnés dans chaque strate,

mais ignorons le nombre d'expéditions dans un comté en

particulier; cette valeur est aléatoire, puisque le nombre

d'expéditions variera selon l'échantillon de chasseurs. Il y a

deux paramètres principaux à estimer et une variable aléa-

toire à prévoir. Le premier paramètre correspond au nombre

total d'expéditions de tous les chasseurs, soit la « pression

de chasse ». Cet aspect importe aux gestionnaires de la

faune qui se préoccupent de la qualité de l'expérience de

chasse. Trop de chasseurs dans une région ont tendance à se

mêler de la chasse d'autrui, d'où une perte de qualité de

cette activité. L'autre paramètre à estimer correspond au

taux de succès de la chasse ou à la proportion de dindons

récoltés par expédition. S'il n'y a pas assez de dindons dans

2.2 Modèle

et Sun (1998), He et Sun (2000) et Oleson et He (2004).

Nous allons présenter un modèle d'estimation du taux de succès et de la pression de la chasse et prévoir simultanément la récolte totale de dindons. Le modèle tient compte de

prévoir le nombre total de dindons récoltés au niveau des comtés et à comparer cette valeur à la valeur de démon-

strées et à prévoir simultanément la récolte totale de dindons. Le modèle tient compte de

prévoir le nombre total de dindons récoltés. On a jugé de

contrôle où on comptait les dindons récoltés. En 1998, les

chasseurs étaient tenus de se rendre à un poste de contrôle

où on comptait les dindons récoltés. On a jugé de contrôler

où on comptait les dindons récoltés. On a jugé de contrôler

où on comptait les dindons récoltés. On a jugé de contrôler

où on comptait les dindons récoltés. On a jugé de contrôler

où on comptait les dindons récoltés. On a jugé de contrôler

modélisons la récolte de dindons à l'aide de distributions

binomiales indépendantes

$$(Y^{ijk} | n^{ijk}, P^{ijk}) \sim \text{Binomial}(n^{ijk}, P^{ijk}), \quad (1)$$

où $i = 1, \dots, I$ est le comté (domaine d'étude), $j = 1, \dots, J$ la période et $k = 1, \dots, K$ la strate nominale ou de

plan d'échantillonnage. Dans ce cas, $I = 114$, $J = 4$, $K = 5$

pour la MTHS. Dans les analyses a priori de cette

enquête, on a posé l'existence d'une taille d'échantillon fixe

n^{ijk} , qui, à notre avis, serait considérée au mieux comme

aléatoire. Nous ignorons le nombre d'expéditions dans

chaque comté tant que les questionnaires n'ont pas été

renvoyés. Nous ignorons également le nombre total d'expé-

ditions dans la petite région N^{ijk} et connaissons seulement

le nombre de chasseurs possible. Comme les chasseurs

doivent s'abstenir après avoir récolté un second dindon, les

comités aux taux de succès supérieurs auraient moins d'ex-

péditions (ou de jours de chasse) pour le même nombre de

chasseurs. S'il y a corrélation, la taille d'échantillon

doit être jugée aléatoire. Ajoutons que, dans un modèle

hiérarchisé de Bayes, si la distribution d'une taille d'échan-

tilion est indépendante de la distribution de la variable de

réponse, les estimations sont identiques pour une taille

d'échantillon (fixe) aléatoire ou non, n , (voir Durbin 1969).

Le taux de succès estimé est plus lisse pour un n^{ijk} fixe que

pour une valeur aléatoire (Woodard, He et Sun 2003).

Maléc et coll. (1997) ont appliqué les estimations régionales

bayésiennes à la National Health Interview Survey (NHIS). Il

existe deux grandes différences entre notre modèle et celui

de Malec et coll. (1997). D'abord, les tailles de population N^{ijk} sont

connues dans le modèle de ces auteurs, mais inconnues dans le nôtre. C'est ce qui explique principalement que nous introduisons un MMLG à deux variables. En second lieu, on modélise dans leur modèle les

taux de succès en logits comme fonction linéaire des covariables. Ainsi, les estimations dépendent des valeurs des

covariables, mais non de la répartition spatiale. Nous ajoutons l'aspect spatial aux covariables dans le calcul

logistique des taux de succès, de sorte que les estimations dépendent à la fois des covariables et de la répartition

spatiale. Cela s'avère d'autant plus nécessaire que des covariables importantes peuvent ne pas être disponibles.

Pour tenir compte du caractère aléatoire des tailles d'échantillon, nous modélisons n^{ijk} par distribution de

Poisson

$$(n^{ijk} | N^{ijk}) \sim \text{Poisson}(R^k N^{ijk}), \quad (2)$$

Pour n^{ijk} , la moyenne et la variance de la distribution de Poisson sont une constante multipliée par la taille de

population N^{ijk} . Cette constante R^k est le rapport entre le nombre échantillonné et le nombre total de chasseurs de la

strate k . On peut la calculer à l'aide du tableau 1.

une de ces petites régions, nous calculons un taux de succès par comté pour chacune des strates nominales et en prenons la moyenne pour une estimation unique de comté. Pour pouvoir combiner les strates du plan d'échantillonnage, nous devrions connaître les tailles individuelles d'effectif des strates. Si ces valeurs sont inconnues, leur estimation est possible par le modèle que nous proposons. Notre propos est donc d'appliquer les méthodes bayésiennes à l'estimation des taux de succès de la chasse aux dindeons au niveau des comtés du Missourï. Nous proposons une nouvelle famille de modèles mixtes linéaires généralisés (MMiLG) à effets aléatoires corrélés lorsque plusieurs paramètres sont inconnus. Nous appellerons notre modèle MMiLG à deux variables. On se sert souvent d'un MMiLG à effets aléatoires corrélés lorsque lorsqu'il n'y a qu'un seul paramètre inconnu (Sun, Speckman et Tsaukakawa 2000). Dans le modèle que nous proposons, nous estimons à la fois la taille de population et le paramètre d'intérêt et, de ce fait, nous faisons progresser la recherche sur les estimations régionales.

l'enquête de 1998 sur la chasse aux dindons au Missouri (« Missourit Turkey Hunting Survey » ou MTHS). Il s'agit d'une enquête postale après la saison printanière qui a renseigné les services de conservation du Missouri sur le nombre de dindons tués par les chasseurs de jour en jour en saison de chasse et sur les comtés où cette récolte a eu lieu. On a aussi relevé dans cette enquête le nombre total de visites de comités par les chasseurs à chaque journée de chasse. On a ensuite calculé les taux de succès de la chasse à l'aide de ces données.

Le cas MTHS est présenté en détail à la section 2. À la section 3, on trouvera un résumé du cadre méthodologique proposé et les formules générales applicables. Le but était d'estimer les taux de succès, mais ces méthodes sont généralisables. Nous offrons aussi des formules générales d'estimation et de prévision pour de petites régions, ainsi que les distributions conditionnelles intégrales pour des simulations MCCM. À la section 4 enfin, nous livrons nos derniers commentaires.

2. Enquête de 1998 sur la chasse aux dindons au Missouri

2.1 Contexte de la MTHS de 1998

Les services de conservation du Missouri ont commencé en 1986 relever à intervalles bienaux les tendances de la chasse aux dindons à l'aide de la MTHS. Dans cette enquête, on a demandé aux chasseurs dans quels comtés ils chassaient, quels étaient leurs jours de chasse et si celle-ci dominait des résultats ou non. Le tout a commencé par un échantillonage aléatoire simple de tous les titulaires de

permis de chasse primaire aux dindons. He et Sun (1998) se sont reportés aux données de l'enquête de 1996 pour estimer les taux de réussite de cette chasse dans les 14 comtés du Missouri en se servant d'un modèle binaire binomial de Bayes. Dans He et Sun (2000), on estime les taux de succès par comté et par semaine de saison de chasse. Pendant la saison primaire de 1996, on n'avait la permission de récolter qu'un dindon par semaine. Ces auteurs ont utilisé un MMLG et estimé les seuls taux de succès. Oleson et He (2004) ont étendu l'application de ce modèle à l'estimation des taux pour chaque jour de la saison de chasse. Ils ont pu voir un taux significatif d'autocorrélation entre les jours de la saison et entre les comtés du Missouri lorsqu'ils ont estimé ces taux de succès.

En 1998, on a modifié le plan d'échantillonnage MTHS. La base de sondage demeurait la liste de tous les chasseurs de dindons ayant la permission de chasser au Missouri. On y trouve entre autres des indications sur le comté de résidence de chaque chasseur. Dans les échantillons aléatoires simples exploités par le passé, les régions populaires de Kansas City et de St. Louis étaient d'un trop grand poids. C'est pourquoi les comtés proches de ces régions métropolitaines se sont vu attribuer un échantillon important et les comtés plus éloignés (plus au sud, par exemple) ont eu droit à un échantillon insuffisant. Idéalement, nous aimerions stratifier selon les comtés où les chasseurs se sont livrés à cette activité afin de pouvoir constituer des échantillons représentatifs à ce niveau des domaines d'étude. Il est exclu de savoir où les intéressés ont chassé avant leur retour des questionnaires, d'où l'impossibilité d'une telle stratification. Une autre possibilité est de stratifier selon les comtés de résidence des chasseurs, puisque ceux-ci ont tendance à chasser près de l'endroit où ils demeurent (c'est-à-dire en des lieux qui leur sont familiers). Cela pose un problème d'estimation des paramètres d'intérêt, c'est-à-dire des taux de succès de la chasse par comté. Nous voudrions des estimations selon les lieux de chasse, mais le plan d'échantillonnage fait intervenir les lieux de résidence des chasseurs. Le nouveau plan de sondage de la MTHS prévoit un échantillonnage aléatoire simple à grappes stratifiées de taille inégale. Dans ce cas, une grappe représente un chasseur titulaire de permis et ses éléments sont les expéditions de chasse de chaque chasseur. Le lieu de résidence du chasseur sert de facteur de stratification. Voici les strates nominales en question : 1) non-résidents du Missouri; 2) résidents du nord du Missouri; 3) résidents du sud du Missouri; 4) résidents de la région métropolitaine de St. Louis; 5) résidents de la région métropolitaine de Kansas City. La figure 1 indique la délimitation des quatre strates de résidents du Missouri en fonction des trois premiers chiffres du code postal américain. On a procédé par répartition proportionnelle pour établir le nombre de chasseurs échantillonnés

Estimation bayésienne pour de petites régions en cas de différence entre les strates du plan d'échantillonnage et les domaines d'étude

Jacob J. Oleson, Chong Z. He, Dongchu Sun et Steven L. Sheriff¹

Résumé

Il s'agit d'obtenir des estimations fiables pour des domaines d'étude où les tailles d'échantillon peuvent être des plus modestes et pour lesquels la strate du plan d'échantillonnage ne coïncide pas avec le domaine. On ignore les tailles de population avant pour le domaine d'étude que pour la strate du plan d'échantillonnage. Dans le calcul des estimations paramétriques des domaines d'étude, le choix d'une taille d'échantillon aléatoire s'impose souvent. Nous proposons une nouvelle famille de modèles mixtes linéaires généralisés (MLMG) à effets aléatoires corrélés lorsqu'il y a plus d'un paramètre inconnu. Le modèle que nous proposons estimera tant la taille de population que le paramètre d'intérêt. Pour ce cadre, nous donnons des formules générales pour les distributions conditionnelles intégrales qui exigent des simulations de Monte Carlo à chaîne de Markov (MCMC). Nous présentons aussi des équations de prévision et d'estimation bayésiennes pour les domaines d'étude. Nous nous servons enfin de l'enquête de 1998 sur la chasse aux dindons dans le Missouri, laquelle stratifie des échantillons en fonction du lieu de résidence du chasseur, et nous voulons obtenir des estimations au niveau du domaine, c'est-à-dire du comté où le chasseur de dindons s'adonne effectivement à cette activité.

Mots clés : Méthode hiérarchisée bayésienne; méthode de Monte Carlo à chaîne de Markov; prévision bayésienne; taille d'échantillon aléatoire; corrélation spatiale; stratification.

1. Introduction

On a souvent affaire à de petites tailles d'échantillon lorsqu'on analyse des données d'enquête. C'est que, dans bien des cas, on se trouve à étudier des sous-populations comme des groupes sociodémographiques. Il se peut aussi que nous considérons des régions dans l'espace ou des périodes dans le temps comme sous-populations ou domaines d'étude. En raison de la faible taille d'échantillon, les estimateurs directs d'enquête pourraient se révéler très peu fiables. On a qualifié d'estimations régionales (ER) les estimations relatives à petite taille d'échantillon. Rao (2003) passe efficacement en revue un grand nombre de techniques ER. On peut aussi trouver un certain nombre d'études récentes sur cette estimation de petites régions dans Rao (2005) et Jiang et Lahiri (2006). On doit disposer de bons modèles pour produire des statistiques fiables sur de petites régions. En modélisation, il existe différentes méthodes, dont celle des meilleures prévisions empiriques (voir Prasad et Rao 1990; Jiang, Lahiri et Wan 2002; Das, Jiang et Rao 2004; Jiang et Lahiri 2006) et les méthodes bayésiennes (voir Mallick, Sedransk, Mortality et LeClerc 1997; Ghosh, Natarajan, Stroud et Carlin 1998; He et Sun 2000). Pour un bon examen des estimations régionales bayésiennes, nous renvoyons le lecteur à Rao (2003), qui traite d'une application bien concrète du cadre méthodologique de Bayes. Une étape primordiale dans l'application d'un modèle bayésien est le choix de distributions a priori. Il faut en outre s'attacher avec soin à la qualité de la

distribution à posteriori et à la robustesse des distributions a priori. Là où une simulation MCMC comme la formule d'échantillonnage de Gibbs entre dans les calculs, on doit contrôler la convergence de la chaîne de Gibbs. Pour plus de détails, voir Carlin et Louis (2000). Regardons un plan d'échantillonnage à grappes stratifiées où chaque grappe est choisie par échantillonnage aléatoire simple sans remise. Dans notre application, les grappes sont d'une taille inégale et les valeurs de taille sont inconnues au moment de concevoir l'enquête. Considérons maintenant un problème d'estimation de domaines où ceux-ci débordent les limites des grappes nominales. On ignore donc la taille de population des domaines et la taille d'échantillon est aléatoire. Dans notre application, les tailles d'échantillon réalisées pour les domaines sont modestes et, en soi, les techniques d'estimation types en fonction du plan d'échantillonnage (voir Cochran 1977, et Lohr 1999) sont peu sûres. Nous proposons comme moyen de contourner ce problème un modèle hiérarchisé intégral de Bayes.

Commençons par obtenir des estimations de taux de succès avec les tailles de population dans de petites régions pour les personnes composant chacune des strates nominales. On obtient ces estimations en tirant de l'information des petites régions voisines, et ce, par une structure spatiale imbriquée dans le modèle bayésien. Ainsi, les estimations obtenues sont bien plus stables que celles de méthodes directes d'enquête. Nous établissons ensuite une moyenne pondérée des taux de succès des strates nominales pour les estimations finales de petites régions. Si un comté constitue

1. Jacob J. Oleson, Département de biostatistique C22-GH, Université de l'Iowa, Iowa City, Iowa, 52242-1009, États-Unis; Chong Z. He et Dongchu Sun, Département de statistique, 146, Middlebush Hall, Université du Missouri à Columbia, Missouri, 65211-6100, États-Unis; Steven L. Sheriff, Ressource Science Center, Département de la conservation du Missouri, 1110, avenue South College, Columbia, Missouri, 65201.

7. Sommaire

D'après les résultats des simulations, nous avons considéré le BRSR et le BRAR comme étant tous deux fiablement exacts sous échantillonnage répété. Conditionnellement à l'échantillon, le BRSR s'est avéré significativement plus efficace (jusqu'à 50 %) que le BRAR pour l'échantillonnage stratifié quand il arrive que la taille d'échantillon de strate soit faible. Par conséquent, le BRSR est la méthode qui a été implémentée dans ABSEST.

Bibliographie

- Brewer, K.R.W., Gross, W.F. et Lee, G.F. (1999). PRN sampling: The Australian experience. *Proceedings of the International Associations of Survey Statisticians*, Helsinki.
- Estévez, V., Hidiroglou, M.A. et Särndal, C.-E. (1995). Methodological principles for a generalised estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 2, 181-204.
- Kovar, J., Rao, J.N.K. et Wu, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 25-44.
- Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 4, 363-378.
- Tam, S.M. (1985). On covariance in finite population sampling. *The Statistician*, 34, 429-433.
- Yung, W., et Rao, J.N.K. (1996). Linéarisation des estimateurs de variance jackknife dans un échantillonnage stratifié à degrés multiples. *Techniques d'enquête*, 22, 23-31.
- Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.
- Singh, A.C., et Mohl, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.
- Recueil disponible au www.fscm.gov.
- ECSM Research Conference, Arlington, VA, 14-16 novembre, 2001. *Presented at 2001 ECSM Research Conference, Arlington, VA, 14-16 novembre, 2001*.
- Roberts, G., Kováčević, M., Mantel, H. et Phillips, O. (2001). Cross sectional inference based on longitudinal surveys: Some experience with Statistics Canada Surveys. *Presented at 2001 ECSM Research Conference, Arlington, VA, 14-16 novembre, 2001*.
- Rao, J.N.K., et Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, mars, 83, 401, 231-241.
- Rao, J.N.K., et Wu, C.F.J. (1984). Bootstrap inference for sample surveys. *Proceeding Section on Survey Methods Research, Journal of the American Statistical Association*, 106-112.
- Preston, J., et Chipperfield, J.O. (2002). Using a generalised estimation methodology for ABS business surveys. *Methodology Advisory Committee* (disponible au www.abs.gov.au).

e) calculer $\hat{\Delta}_{reg}^* = \hat{Y}_{reg}^{*(2)} - \hat{Y}_{reg}^{*(1)}$ ou $\hat{Y}_{reg}^{*(i)} = \sum_{j \in \mathcal{S}^{(i)}} w_{hj}^{*(i)} Y_j^*$. L'estimateur BRSR de la variance est donné par

$$\widehat{\text{Var}}_B(\hat{\Delta}_{reg}^*) = (R-1)^{-1} \sum_{r=1}^R (\hat{\Delta}_{reg}^* - \bar{\Delta}_{reg}^*)^2,$$

ou

$$\hat{\Delta}_{reg}^{(1)} = \hat{Y}_{reg}^{(1)} - \hat{Y}_{reg}^{(2)}$$

et

$$\hat{Y}_{reg}^{(i)} = \sum_{j \in \mathcal{S}^{(i)}} w_{hj}^{*(i)} Y_j^*.$$

La preuve que $\widehat{\text{Var}}_B(\hat{\Delta}_{reg}^*)$ est sans biais est simple et semblable à la preuve que $\text{Var}_B(\hat{\theta})$ est sans biais (voir la section 4).

L'approche décrite ci-dessus requiert un ensemble distinct de poids de rééchantillonnage pour les estimations de la variance des changements et des niveaux. Roberts, Kovacevich, Mantel et Phillips (2001) considèrent des estimateurs bootstrap approximatifs de la variance des changements utilisant uniquement les poids de rééchantillonnage pour les niveaux, ce qui réduit les coûts de calcul et simplifie la méthode et sa mise en œuvre dans un système informatique.

6. Étude par simulation

À la présente section, nous résumons une étude par simulation réalisée pour les estimations ponctuelles et les estimations des changements afin de mesurer empiriquement le biais et la variabilité du BRSR et du BRAR sous un échantillonnage répété quand $R = 100$. Nous avons généré une population aux points 1 et 2 dans le temps à partir des modèles suivants, $Y_j^{(1)} = (0,75x_{1j}^{(1)} + 0,25x_j^{(1)})W(0; 2,5; 1)$ et $Y_j^{(2)} = (1,5Y_j^{(1)} + 0,5Y_j^{(1)})W(0; 5; 1)$, où les variables auxiliaires sont données par $x_{1j} = 0,25x_j + 0,75[100I(0; 1; 1)]$ et $x_j = 100I(0; 1; 1)$ ou $W(\mu; \gamma; \alpha)$ et $L(\mu; \gamma; \alpha)$ sont les lois de Weibull et log-normale dont les paramètres d'emplacement, de forme et de dispersion sont donnés par μ, γ et α . Ces lois reflètent les longues queues des distributions typiques des données d'enquête économique. Les tailles de population aux temps 1 et 2 étaient de 3 000, avec 2 500 unités communes aux deux points dans le temps. Chaque unité de population, i , a été assignée à l'une des cinq strates aux deux points dans le temps en utilisant z_i , où $z_i = x_{1i}W(0; 2,5; 1)$ et les bornes des strates étaient $z_i = 50, 100, 150, 250$. Nous avons ainsi obtenu des tailles de population de strate variant de 400 à 1 000.

Nous avons tiré, en tout, 3 000 EASSR stratifiés simulés à partir de la population aux temps 1 et 2, où $n^{(i)} = 12$, $n^{ch} = n^{(2)} = 4$ et $n^{ch} = 8$ pour tout h et $i = 1, 2$. Pour le BRSR, les tailles des échantillons répétées sont données aux sections 4 et 5. Pour le BRAR, les tailles des échantillons

L'erreur-type de \hat{Y}_{reg}^* estimée par le bootstrap à partir du j^{e} échantillon est

$$S_{j^*} = \sqrt{\frac{1}{100} \sum_{r=1}^{100} (\hat{Y}_{j^*r}^{reg} - \bar{Y}_{j^*}^{reg})^2},$$

où $\hat{Y}_{j^*r}^{reg}$ est défini de façon analogue à $\hat{Y}_{j^*}^{reg}$. Le biais relatif (BR) de l'erreur-type estimée par le bootstrap est

$$\text{BR}(S) = \frac{3000}{1} \sum_{j=1}^{3000} (S_{j^*} - S).$$

La racine carrée relative de l'erreur quadratique moyenne (RREQM) de l'erreur-type estimée par le bootstrap est

$$\text{RREQM}(S) = \frac{1}{S} \sqrt{\frac{1}{3000} \sum_{j=1}^{3000} (S_{j^*}^2 - S^2)}.$$

Des définitions semblables de la RREQM et du biais sont utilisées pour l'estimation de la variance des changements. Nous comparons aussi les probabilités de couverture à 95 %, c'est-à-dire le pourcentage des intervalles de confiance à 95 % contenant le total de population réel du BRSR et du BRAR pour les niveaux et les changements. Les résultats présentés dans le tableau 1 montrent que le BR et la RREQM du BRSR et du BRAR sont tous deux acceptablement faibles. Le biais des estimations ponctuelles au temps 1 produit par le BRAR est légèrement plus grand que celui obtenu pour le BRSR, ce qui se traduit par des probabilités de couverture un peu moins bonnes.

Tableau 1
Estimation bootstrap de l'erreur-type des changements et des estimations ponctuelles

Temps 1		Changement	
Méthode	BR	RREQM	C 95 %
BRSR	0,7	17,3	94,7
BRAR	-3,1	15,8	93,7
BRSR	0,7	17,3	94,7
BRAR	-3,1	15,8	93,7
C 95 %	95,3	20,7	2,1
RREQM	95,3	19,6	1,3

5. Variance du changement entre des estimations à une seule phase

Un important produit attendu de nombreuses enquêtes-

entreprises est l'estimation du changement entre deux points dans le temps. Représentons la population finie au temps t par $U^{(t)}$, où $U^{(t)}$ est la population de la strate h au temps t qui est constituée de $N^{(t)}$ unités. Le total de population au temps t est $Y^{(t)} = \sum^h \sum_{i \in U^{(t)}} y_{hi}^{(t)}$. La présente section porte sur l'estimation de la variance de $\Delta^{(t)} = Y^{(t)} - Y^{(t-1)}$, c'est-à-dire la différence entre deux périodes. Les termes correspondant à $n^{(t)} f^h$ et s^h au temps t sont dénotés par $n^{(t)} f^{(t)}$ et $s^{(t)}$, respectivement. En cas d'échantillonnage à deux occasions, définissons $N^{(t)} n^{(t)}$ et $n^{(t)}$ comme étant le nombre d'unités dans les ensembles suivants : $s^{(t)} - s^{(t-1)} = s^{(t)}_{(2)} - s^{(t-1)}_{(2)}$, respectivement. Dans les enquêtes-entreprises de l'ABS, l'échantillon au temps 1 de taille $n^{(1)}$ est un EASSR tiré de $U^{(1)}$. L'échantillon au temps 2 est l'union des deux échantillons suivants : un EASSR de $n^{(2)}$ unités tirées de $s^{(t)}$ (et un EASSR de $n^{(2)}$ unités tirées de $U^{(1)} \cap U^{(2)}$). L'échantillon au temps 2 est, en fait, un EASSR tiré de $U^{(2)}$. À l'ABS, la taille de l'échantillon chevauchant, $n^{(t)}$, est contrôlée par la méthode du nombre aléatoire permanent (voir Brewer, Gross et Lee 1999).

L'estimateur de $\text{Var}(\Delta)$ peut être exprimé par
$$\text{Var}(\Delta) = \text{Var}(Y^{(1)} + \text{Var}(Y^{(2)}) - 2\text{Cov}(Y^{(1)}, Y^{(2)}).$$
 Considérons l'estimateur d'Horvitz-Thompson $\Delta = Y^{(2)} - Y^{(1)}$, où $t = 1, 2$ et Y^t est défini de manière analogue à Y^t . Tam (1985) montre que, si $U^{(t)}$ est un estimateur sans biais de $\text{Var}(\Delta)$ sous le plan d'échantillonnage susmentionné est donné par

où
$$\widehat{\text{Var}}(Y^{(t)}) = \sum^h N^h_{(2)} (1 - f^{(t)}_h) s^{(t)}_{(2)} / n^{(t)}_{(2)} \quad \widehat{\text{Cov}}(Y^{(1)}, Y^{(2)}) = \sum^h N^h_{(2)} (1 - f^{(1)}_h) s^{(1)}_{(2)} n^{(2)}_{(2)} / (n^{(1)}_{(2)} n^{(2)}_{(2)})$$

$$\hat{y}^t_i = \sum_{j \in s^{(t)}_{(2)}} y_{ji} = n^{(t)}_{(2)} / n^{(t)}_{(2)} \quad \text{Si } U^{(1)}_{(2)} \neq U^{(2)}_{(2)}, \text{ une forme plus générale de l'estimateur de Tam est donnée par } \widehat{\text{Var}}(\Delta), \text{ excepté que}$$

$$\widehat{\text{Var}}(Y^{(t)}) = \sum^h N^{(t)}_{(2)} (1 - f^{(t)}_h) s^{(t)}_{(2)} / n^{(t)}_{(2)} \quad \widehat{\text{Cov}}(Y^{(1)}, Y^{(2)}) = \sum^h N^{(1)}_{(2)} N^{(2)}_{(2)} / (n^{(1)}_{(2)} n^{(2)}_{(2)}) n^{(1)}_{(2)} (1 - f^{(1)}_h) s^{(1)}_{(2)}$$

et

$$f^{(1)}_{12,h} = \frac{n^{(1)}_{(2)} N^{(2)}_{(2)}}{n^{(1)}_{(2)} N^{(2)}_{(2)}}.$$

Dans la suite de la présente section, nous supposons que $\widehat{\text{Var}}(\Delta)$ est sans biais pour $\text{Var}(\Delta)$ quand $U^{(t)} \neq U^{(t-1)}$. (II) convient de souligner que $\widehat{\text{Var}}(\Delta)$ peut prendre des valeurs négatives quand $U^{(1)}_{(2)} \neq U^{(2)}_{(2)}$. Nordberg (2000) donne un estimateur sans biais de $\text{Var}(\Delta)$ pour l'estimateur par la régression quand $U^{(1)}_{(2)} \neq U^{(2)}_{(2)}$, mais il n'existe aucun moyen évident de l'utiliser avec la méthode bootstrap décrite dans le présent article.) L'estimation de la variance de $\Delta^{(t)} = Y^{(t)} - Y^{(t-1)}$ le changement entre les estimations GREG aux temps 1 et 2, en utilisant le BRSR requiert de répéter R fois les étapes suivantes :

- a) former l'ensemble s^* en sélectionnant indépendamment $m^{ch}_{(1)} = [n^{ch}_{(1)} / 2]$ et $m^{ch}_{(2)} = [n^{ch}_{(2)} / 2]$ unités par EASSR à partir des ensembles $s^{(1)}_{hc}$ et $s^{(2)}_{hc}$, respectivement;
- b) pour $t \in s^{(1)}_{hc}$, calculer les poids de rééchantillonnage
$$w^{hi}_{(1)} = N / n^{hi}_{(1)} \left[1 - \gamma^{ch}_{(1)} \frac{m^{hi}_{(1)}}{m^{ch}_{(1)}} + \gamma^{ch}_{(1)} \frac{m^{hi}_{(1)}}{m^{ch}_{(1)}} \right]$$
 pour $t \in s^{(2)}_{hc}$
$$w^{hi}_{(2)} = \left[1 - \gamma^{ch}_{(2)} \frac{m^{hi}_{(2)}}{m^{ch}_{(2)}} + \gamma^{ch}_{(2)} \frac{m^{hi}_{(2)}}{m^{ch}_{(2)}} \right] \text{ ou } \gamma^{ch}_{(2)} = \sqrt{(1 - f^{(2)}_{12,h}) m^{ch}_{(2)} / (n^{ch}_{(2)} - m^{ch}_{(2)})}$$
 et $\gamma^{ch}_{(1)}$ égale 1 si l'unité t est sélectionnée dans le groupe échantillonné au temps 1 et 0 autrement;
- c) calculer les poids définis de façon analogue pour $t \in s^{(2)}_{hc}$;
- d) calculer $w^{hi}_{(t)} = w^{hi}_{(1)} s^{(1)}_{(2)}$ pour $t \in s^{(1)}_{hc}$ et $w^{hi}_{(t)} = w^{hi}_{(2)} s^{(2)}_{(2)}$ pour $t \in s^{(2)}_{hc}$, mais est calculé en utilisant les poids $w^{hi}_{(t)}$ au lieu de $w^{hi}_{(t)}$;

4.2 Note sur l'efficacité relative des échantillonnages BRAR et BRSR

Afin de simplifier la notation, posons que $\hat{V}^{\text{boot}} = \widehat{\text{Var}}_B(\hat{\theta})$. La variance de l'estimateur bootstrap de la variance peut s'écrire

$$\widehat{\text{Var}}(\hat{V}^{\text{boot}}) = \text{Var}_s(E_s[\hat{V}^{\text{boot}} | s]) + E_s[\widehat{\text{Var}}_B(\hat{V}^{\text{boot}} | s)],$$

où s désigne l'espérance par rapport au plan d'échantillonnage. Si \hat{V}^{boot} est sans biais (c'est-à-dire $E_s[\hat{V}^{\text{boot}} | s] = \text{Var}(\hat{\theta})$), alors $\widehat{\text{Var}}(\hat{V}^{\text{boot}})$ ne dépend pas de la façon de sélectionner les échantillons répétés. Le terme $\text{Var}_s(\hat{V}^{\text{boot}} | s)$ est l'erreur de rééchantillonnage sachant l'échantillon et est inversement proportionnel à R . La valeur de R est choisie suffisamment grande pour que $\text{Var}_s(E_s[\hat{V}^{\text{boot}} | s])$ soit faible comparativement à \hat{V}^{boot} , la variance d'échantillon estimée.

L'efficacité des deux estimateurs bootstrap peut être comparée en se basant sur la grandeur de $\text{Var}_s(E_s[\hat{V}^{\text{boot}} | s])$ quand les deux estimateurs ont la même valeur de R . Nous allons maintenant résumer les résultats empiriques basés sur des données réelles qui montrent que le BRSR peut être significativement plus efficace que le BRAR. Les gains d'efficacité se traduisent par une réduction du temps de calcul et/ou l'obtention d'estimations plus exactes de la variance.

Preston et Chipperfield (2002) ont comparé l'efficacité du BRSR avec $m_h = [n_h/2]$ et du BRAR avec $m_h = n_h - 1$ (voir Rao et Wu 1984) dans le cas de l'enquête trimestrielle sur l'activité économique en Australie de mars 2000. La taille de l'échantillon de cette enquête au niveau de la strate varie de quatre à plusieurs centaines. Les résultats (tirés de Preston et Chipperfield 2002, tableau 1) montrent qu'au niveau national, $\text{Var}_s(E_s[\hat{V}^{\text{boot}} | s])$ est 54 % plus faible pour l'échantillonnage BRSR que pour l'échantillonnage BRAR quand $R = 100$ (Voir Preston et Chipperfield 2002, pour plus d'estimations empiriques de $\text{Var}_s(E_s[\hat{V}^{\text{boot}} | s])$ pour le BRAR et le BRSR). Autrement dit, le BRSR nécessite environ deux fois moins de répétitions que le BRAR pour atteindre la même erreur de rééchantillonnage, ce qui représente un gain d'efficacité important. Un autre avantage du BRSR par rapport au BRAR est que le temps machine pour la sélection des échantillons répétés est considérablement plus court.

Dans le cas des études empiriques, le choix de $m_h = [n_h/2]$ pour le BRSR minimise $\text{Var}_s(E_s[\hat{V}^{\text{boot}} | s])$. Nous soupçonnons qu'à mesure que n augmente, l'écart entre le BRSR et le BRAR tendra à devenir nul. D'autres travaux sont nécessaires en vue d'établir ces propriétés.

b) calculer $w_{hi}^* = w_{hi}(1 - \gamma_h + \gamma_h m_h / m_h \hat{\delta}_{hi}^*)$ pour $i \in s_h$, où $\gamma_h = \sqrt{(1 - f_h) m_h / (n_h - m_h)}$, $\hat{\delta}_{hi}^*$ est égal à 1 si $i \in s_h^*$ et à 0 autrement;

c) calculer $w_{hi}^* g_{hi}^* = w_{hi}^* g_{hi}^*$ pour $i \in s$, et

d) calculer la r^{e} estimation bootstrap de $\sum_{i \in s} w_{hi}^* g_{hi}^*$. La justification de $m_h = [n_h/2]$ est présentée à la section 4.2. L'estimateur de variance bootstrap est donné par l'approximation de Monte Carlo, $\widehat{\text{Var}}_B(\hat{Y}^{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^R (\hat{Y}_r^{\text{reg}} - \bar{Y}^{\text{reg}})^2$. La méthode BRAR est identique à la méthode BRSR, excepté que les échantillons répétés sont sélectionnés par BASAR et que le facteur d'échelle devient $\gamma_h = \sqrt{(1 - f_h) m_h / (n_h - 1)}$, où m_h est souvent fixé à $n_h - 1$ dans la littérature. Preston et Chipperfield (2002) ont constaté que l'erreur de rééchantillonnage, c'est-à-dire l'erreur due à l'échantillonnage répété et au conditionnement plus faible pour le BRSR que pour le BRAR.

Il est facile de voir que les estimateurs BRSR et BRAR sont des estimateurs sans biais de $\text{Var}(\hat{\theta})$. La variance approximative par développement en série de Taylor est donnée par $\widehat{\text{Var}}(\hat{\theta}) = \nabla' \hat{V}(\hat{Y}) \nabla \hat{\theta}$, où $\hat{V}(\hat{Y})$ est la matrice de dimensions $P \times P$ contenant les éléments

$$\widehat{\text{Cov}}(\hat{Y}^p, \hat{Y}^p) = \frac{n}{N^2(1 - f)} \hat{s}_{p,p},$$

où

$$\hat{s}_{p,p} = \frac{1}{n-1} \sum_{i \in s} (\hat{Y}^{pi} - \bar{\hat{Y}}^p)(\hat{Y}^{pi} - \bar{\hat{Y}}^p);$$

$$\bar{\hat{Y}}^p = \frac{1}{n} \sum_{i \in s} \hat{Y}^{pi};$$

$$\hat{Y}^p = \sum_{i \in s} w_i \hat{Y}^{pi}$$

pour $p, p' = 1, \dots, P$, et $\nabla' = (\partial/\partial Y_1, \dots, \partial/\partial Y_P)'$. Il est aisé de constater que

$$E_s(\widehat{\text{Var}}(\hat{\theta})) = \nabla' \hat{V}_s(E_s(\hat{Y}^p)) \nabla \hat{\theta} = \nabla' \hat{V}(\hat{Y}) \nabla \hat{\theta},$$

en notant que

$$E_s[\widehat{\text{Cov}}(\hat{Y}^p, \hat{Y}^{p'})] = \widehat{\text{Cov}}(\hat{Y}^p, \hat{Y}^{p'})$$

où E_s désigne l'espérance sous rééchantillonnage. Notons que les constantes d'échelle appliquées à w_{hi} pour calculer les poids de rééchantillonnage sont choisies de façon à obtenir le facteur de correction de population finie correct. Il s'ensuit que l'approximation de Monte Carlo de la variance, $\widehat{\text{Var}}_B(\hat{\theta}) = (R-1)^{-1} \sum_{r=1}^R (\hat{\theta}^{*r} - \bar{\hat{\theta}})^2$, est sans biais pour $\text{Var}(\hat{\theta})$.

3. Comparaison de divers estimateurs de variance

La méthode d'estimation de la variance adoptée pour ABSEST devait avoir, dans les études par simulation, de bonnes propriétés de biais et de variance comparativement aux options décrites dans la littérature. Afin de simplifier la maintenance et le développement du système, il fallait que les spécifications du système d'estimation de la variance soient génériques, afin que tous les calculs soient en grande partie indépendants de l'estimateur. (ABSEST ne doit être capable de prendre en charge que l'EASSR dans les strates et les plans d'échantillonnage à un degré.) Enfin, la réduction au minimum des coûts de calcul était une autre considération importante.

En premier lieu, nous avons également envisagé les méthodes du bootstrap, du jackknife et des répliques équilibrées répétées (BRR pour Balanced Repeated Replication) (Shao et Tu 1995, Rao et Wu 1988). Considérons l'estimation de la variance d'une fonction $\theta = \theta(\bar{Y})$, où \bar{Y} est un vecteur de dimension P d'estimations $Y = \sum_{i \in s} w_i y_i$, y_i éléments y_{pi} et θ est une fonction lisse. L'estimation de la variance par une méthode de rééchantillonnage comporte les étapes suivantes :

- i) sous-échantillonner indépendamment l'ensemble s un nombre total R de fois;
- ii) pour chacun des R sous-échantillons, calculer $w_i^* = b_i^* w_i$, où b_i^* dépend du nombre de fois que l'unité i est sélectionnée dans le sous-échantillon;

- iii) calculer $\hat{\theta}^* = \theta(\bar{Y}^*)$, où $\bar{Y}^* = \sum_{i \in s} w_i^* y_i^*$;
- iv) estimer la variance de $\hat{\theta}$ par $\text{Var}^{\text{rep}}(\hat{\theta}) = (R-1)^{-1} \sum_{r=1}^R (\hat{\theta}^{(r)} - \bar{\theta})^2$, où $\bar{\theta}^{(r)}$ est l'estimation de θ basée sur le r^{e} échantillon répété. Nota : l'expression pour les poids de rééchantillonnage, $w_i^* = b_i^* w_i$, comprend le jackknife, le bootstrap et les répliques répétées équilibrées comme cas particuliers.

Comme \bar{Y}^{reg} peut être exprimé par une fonction θ , nous pouvons calculer la variance de \bar{Y}^{reg} suivant les étapes susmentionnées où iii) et iv) deviennent, respectivement : iii) calculer $\bar{Y}^{\text{reg}} = \sum_{i \in s} w_i^* g_i$, où $w_i^* = w_i g_i^*$ et g_i^* a la même forme que g_i , mais est calculé en utilisant les poids w_i^* au lieu des poids w_i pour $i \in s$; iv) estimer la variance par $\text{Var}^{\text{rep}}(\bar{Y}^{\text{reg}}) = (R-1)^{-1} \sum_{r=1}^R (\bar{Y}^{\text{reg}(r)} - \bar{Y}^{\text{reg}})^2$.

La caractéristique intéressante de ces méthodes de rééchantillonnage est que seules la sélection des échantillons répétées et la valeur b_i^* sont nécessaires pour calculer des estimations sans biais de la variance pour de nombreux plans d'échantillonnage utilisés couramment et pour des estimateurs ayant de bonnes approximations par développement en série de Taylor au premier ordre. En outre, si l'on

le bootstrap rééchantonné sans remise (BRSR) [en anglais, without replacement scaled bootstrap] et par le bootstrap rééchantonné avec remise (BRAR) [en anglais, with replacement scaled bootstrap] de Rao et Wu (1988) pour les estimations ponctuelles sous des plans de sondage à une phase. À la section 5, nous décrivons le BRSR pour les estimations de changements. À la section 6, nous évaluons les propriétés de biais et de variance du BRSR et du BRAR au moyen d'une étude par simulation. À la section 7, nous présentons certaines conclusions.

2. Estimateur par la régression généralisée (GREG)

À la présente section, nous décrivons brièvement l'estimateur GREG qui est implémenté dans ABSEST. Considérons une population finie U divisée en H strates $U = \{U_1, U_2, \dots, U_H\}$, où U_h est constituée de N_h unités. Le total de population finie étudié est $Y = \sum_{h=1}^H Y_h$, où $Y_h = \sum_{i \in U_h} y_{hi}$ et $h = 1, \dots, H$. Dans la strate h , l'échantillon s_h de n_h unités est sélectionné à partir de U_h par échantillonnage aléatoire simple sans remise (EASSR). L'ensemble complet d'échantillons est désigné par $s = \{s_1, s_2, \dots, s_H\}$.

Considérons le cas où un vecteur de variables auxiliaires $x_i = (x_{i1}, \dots, x_{iK})^T$ de dimension K est disponible pour $i \in s$ et le vecteur correspondant de totaux de population $X = \sum_{i \in U} x_i$ est connu. L'estimateur GREG (Särndal, Swensson et Wretman 1992, page 227) est donné par $\bar{Y}^{\text{reg}} = \sum_{i \in s} w_i y_i + (X - X^T) \bar{b}$, où $w_i = w_i g_i$, $w_i = 1/N_h$ avec $\bar{b} = \bar{b}^T$ avec \bar{b}^T représentant l'inverse généralisée de $\bar{F}^T (X - X^T) \bar{F}$, $\bar{F} = \sum_{i \in s} w_i x_i x_i^T$, $\bar{F}^{-1} = \sum_{i \in s} w_i x_i y_i \sigma_i^{-2}$, $\sigma_i^2 = x_i^T \beta + \varepsilon_i$ telle que les résidus ε_i sont indépendants identiquement distribués de moyenne 0 et de variance σ_i^2 , et $E(\bar{b}) = \beta$. Il est bien connu que \bar{Y}^{reg} est sans biais jusqu'à l'ordre $O(n^{-1})$. Les poids w_i , w_i^* donnés par l'équation susmentionnée se situent en dehors de ces bornes, ils sont calculés par itération (voir la méthode 5 de Singh et Mohl 1996).

L'expression de \bar{Y}^{reg} peut être adaptée à une gamme d'estimations, y compris des estimations de domaines et des estimations à plusieurs phases (voir Estevao, Hidiroglou et Särndal 1995). Par exemple, si $x_i = 1$, \bar{Y}^{reg} devient l'estimateur d'Horvitz-Thompson donné par $\bar{Y} = \sum_{h=1}^H n_h \sum_{i \in s_h} y_{hi} / N_h$ avec une variance estimée $\text{Var}(\bar{Y}) = \sum_{h=1}^H N_h n_h^{-1} (1 - f_h) s_{h2}^2 / n_h$ où $s_h^2 = (n_h - 1)^{-1} \sum_{i \in s_h} (y_{hi} - \bar{y}_h)^2$, $\bar{y}_h = \sum_{i \in s_h} y_{hi} / n_h$ et $f_h = n_h / N_h$.

Bootstrap efficace pour les enquêtes-entreprises

James Chipperfield et John Preston¹

Résumé

L'Australian Bureau of Statistics vient de développer un système généralisé d'estimation pour traiter les données de ses enquêtes-entreprises annuelles et infra-annuelles de grande portée. Les plans de sondage de ces enquêtes comportent d'un grand nombre de strates, un échantillonnage aléatoire simple dans les strates, des fractions d'échantillonnage non négligeables, ainsi qu'un chevauchement d'échantillons pour des périodes consécutives et peuvent faire l'objet de modifications de la base de sondage. Un défi important consistait à choisir la méthode d'estimation de la variance répondant le mieux aux critères suivants : être valide pour une grande gamme d'estimateurs (par exemple, ratio et régression généralisée), nécessiter un temps de calcul limité, être facilement adaptable à divers plans de sondage et estimateurs, et avoir de bonnes propriétés théoriques en ce qui concerne le biais et la variance. Le présent article décrit le bootstrap rééchantillonné avec remise (BRAR) de Rao et Wu (1988). Les principaux avantages du bootstrap comparativement à d'autres estimateurs de variance par rééchantillonnage sont son efficacité (c'est-à-dire son exactitude par unité d'espace de mémorisation) et la simplicité relative avec laquelle il peut être spécifié dans un système. Le présent article décrit l'estimateur de variance du bootstrap BRSR pour les estimations ponctuelles et les estimations des changements qui peut être exprimé comme une fonction des moyennes de population finie. Les résultats des simulations entreprises dans le cadre du processus d'évaluation montrent que le BRSR est plus efficace que le BRAR, particulièrement dans les situations où la taille des échantillons dans les strates peut être aussi petite que 5.

Mots clés : Variance; bootstrap; échantillonnage stratifié.

1. Introduction

En 2000, l'Australian Bureau of Statistics (ABS) a commencé par obtenir auprès de l'Australian Taxation Office (ATO) un registre des entreprises contenant les données fiscales. Les éléments de données incluaient le chiffre d'affaires, le chiffre des ventes et d'autres postes de dépenses. En 2001, l'ABS a utilisé ce registre comme base de sondage pour certaines enquêtes afin d'accroître l'efficacité de ses plans d'échantillonnage. Les données du registre sont mises à jour au moins annuellement pour chaque entreprise. Afin de maximiser l'utilisation de ces éléments de données administratives à l'étape de l'estimation, l'ABS a développé un système généralisé d'estimation appelé ABSEST capable de prendre en charge l'estimation par la régression généralisée (GREG) et l'estimation de la variance. Depuis juillet 2005, ABSEST a été utilisé de façon systématique pour l'enquête mensuelle sur le commerce de détail menée par l'ABS.

Un système généralisé d'estimation est un outil fort précieux pour les organismes statistiques, car il permet de gérer diverses exigences quant aux produits d'une enquête en maintenant une grande rigueur statistique à un coût acceptable. L'ABS a investi d'importantes ressources dans son système généralisé d'estimation pour les enquêtes-entreprises. Avant 1998, ce système permettait de produire des estimations de type Horvitz-Thompson, par le ratio et à deux phases, ainsi que des estimations de la variance

fondées sur des approximations par développement en série de Taylor (ST). En 1999, la méthode de développement en série de Taylor a été remplacée par celle du jackknife. Les réactions concernant la conception et la convivialité de l'application informatique ont révélé que les changements apportés au système généralisé d'estimation rendaient sa maintenance et son développement de plus en plus difficiles et que le temps de traitement pouvait être indésirablement long. Ces caractéristiques essentielles ont joué un rôle important dans le choix de la méthode d'estimation de la variance pour ABSEST.

Les données de base des enquêtes-entreprises de l'ABS sont des estimations ponctuelles, des estimations de changements entre deux points dans le temps, et des estimations de taux. Les enquêtes-entreprises sont réalisées selon un plan d'échantillonnage avec probabilités égales dans les strates comportant un très grand nombre de strates (des centaines), à une ou à deux phases et pour les enquêtes comportant un échantillonnage à plus d'une occasion, un chevauchement des échantillons pouvant varier de 0 à 100 %. La taille d'échantillon des enquêtes-entreprises va de moins de 1 000 à 15 000; au niveau des strates, elle peut être aussi faible que trois ou atteindre plusieurs centaines.

À la section 2, nous présentons l'estimateur GREG. À la section 3, nous discutons de divers estimateurs de la variance pour l'estimateur GREG et justifions le choix de l'estimateur de variance par le bootstrap pour ABSEST. À la section 4, nous décrivons les estimateurs de variance par

4.3 Estimation de $E^q V^m E^{pi}(R_I - R | a, b)$ sous l'approche MI

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $E^q V^m E^{pi}(R_I - R | a, b)$ peut être approximé par

$$E^q V^m E^{pi}(R_I - R | a, b) \approx 1 \left[\frac{E^q(N_a)}{1} - 1 \right] \left(\frac{E^q(N_a)}{1} - 1 \right) \sigma^2 + \frac{(\mu_x)^2}{2} \left(\frac{E^q(N_b)}{1} - 1 \right) \left(\frac{E^q(N_b)}{1} - 1 \right) \sigma^2 - 2 \left(\frac{\mu_x}{N_{ab}^b} \right) \left(\frac{E^q(N_{ab}^b)}{1} - 1 \right) \sigma_{en}^2 \right], \quad (4.8)$$

où $N_{ab} = \sum_{i \in U} a_i b_i$. Nous obtenons un estimateur de $E^q V^m E^{pi}(R_I - R | a, b)$ en estimant les quantités in-

$$V^2_{(M)} = \frac{1}{N} \left[\frac{1}{2} \left(\frac{N_a}{1} - 1 \right) \left(\frac{N_a}{1} - 1 \right) \sigma^2 + \frac{N_b}{2} \left(\frac{N_b}{1} - 1 \right) \left(\frac{N_b}{1} - 1 \right) \sigma^2 \right]$$

$$- 2R_I' \left(\frac{N_{ab}^b}{N} - \frac{1}{N} \right) \sigma_{xy}^2 \right]. \quad (4.9)$$

L'estimateur (4.9) est asymptotiquement $mpd1$ - sans biais pour la variance approximative (4.8). Il est intéressant de souligner que, sous l'imputation HDAM pondérée, l'estimateur $V^2_{(NM)}$ dans (4.6) obtenu sous l'approche MN est identique à l'estimateur $V^2_{(M)}$ dans (4.9) obtenu sous l'approche ML. Toutefois, il pourrait ne pas en être ainsi sous une méthode d'imputation différente. En outre, la composante $V^2_{(M)}$ est négligeable par rapport à $V^2_{(NM)}$ quand la fraction d'échantillonnage n/N est négligeable, où $V^2_{(NM)}$ représente $V^2_{(NM)}$ ou $V^2_{(M)}$. Dans ce cas, la composante $V^2_{(NM)}$ peut être omise des calculs.

Enfin, un estimateur de la variance totale sous le cadre inverse est donné par

$$V^{(RE)TOT} = V^1 + V^I + V^2.$$

Sous le cadre inverse, les approches MN et MI mènent toutes deux au même estimateur de la variance totale. Donc, l'estimateur de variance $V^{(RE)TOT}$ est robuste en ce sens qu'il est valide sous l'approche MN ou ML.

5. Résumé et conclusion

Dans le présent article, nous avons obtenu des estimateurs de variance pour l'estimateur imputé d'un ratio sous deux cadres différents. Le cadre inverse facilite la détermination des expressions de la variance (comparativement au cadre à deux phases habituel), particulièrement si la fraction d'échantillonnage est faible, auquel cas nous pouvons omettre la composante $V^2_{(NM)}$. Toutefois, contrairement au cadre à deux phases, il nécessite l'hypothèse supplémentaire voulant que les probabilités de réponse ne dépendent pas de l'échantillon réalisé. De plus, le cadre à deux phases

Bibliographie

Brick, M.J., Kalton, G. et Kim, J.K. (2004). Estimation de variance pour l'imputation hot deck à l'aide d'un modèle. *Techniques d'enquête*, 30, 63-72.

Deville, J.-C., et Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

Rao, J.N.K., et Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Särndal, C.-E. (1992). Méthodes pour estimer la précision des estimations d'une enquête ayant fait l'objet d'une imputation. *Techniques d'enquête*, 18, 257-268.

Shao, J., et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sous l'approche MIN ainsi que sous l'approche ML, nous pouvons obtenir un estimateur du premier terme du deuxième membre de (4.1) et de (4.2) en trouvant un estimateur asymptotiquement pI - sans biais de $V^p E_I(\hat{R}_I | \mathbf{a}, \mathbf{b})$. En outre, nous pouvons estimer le deuxième terme du deuxième membre de (4.1) et de (4.2) par I_1 donné par (3.6). Sous l'approche MN, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.1) en estimant $V^p E_{PI}(\hat{R}_I | \mathbf{a}, \mathbf{b})$, tandis que sous l'approche ML, nous pouvons obtenir un estimateur du dernier terme du deuxième membre de (4.2) en estimant $V^m E_{PI}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$. Par conséquent, les estimateurs de deux premiers termes de (4.1) et de (4.2) sont identiques et donc valides, quelle que soit l'approche (MN ou ML) utilisée pour l'inférence. Seul le troisième terme du deuxième membre de (4.1) et de (4.2) dépend de l'approche choisie. Dans le cas de l'approche ML, la spécification et la validation du modèle d'imputation sont essentielles à l'obtention de l'absence de biais de la troisième composante, tandis que dans l'approche MN, l'absence de biais de la troisième composante dépend de la spécification correcte du modèle de non-réponse.

4.1 Estimation de $V^p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$

Partant d'un développement en série de Taylor du premier ordre et de l'expression (2.5), un estimateur de $V^p E_I(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$, dénoté par I_1^p , est donné par

$$I_1^p = \sum_{i \in S} \sum_{j \in S} \Delta_{ij} \hat{\xi}_i \hat{\xi}_j, \quad (4.3)$$

où

$$\hat{\xi}_i = \frac{1}{N} \left[\frac{1}{N} a_i (y_i - \bar{y}) - \left(\frac{\bar{y}}{N} \right) b_i (x_i - \bar{x}) \right].$$

Autrement dit, l'estimateur I_1^p est obtenu à partir de l'estimateur de variance sous données complètes (2.2) en remplaçant e_i par $\hat{\xi}_i$. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.3) se réduit à

$$I_1^p = \left(1 - \frac{1}{N} \right) \left[\frac{1}{N^2} \sum_{i \in S} \sum_{j \in S} \left(\frac{y_i}{N} - \frac{y_j}{N} \right) \left(\frac{x_i}{N} - \frac{x_j}{N} \right) \right]. \quad (4.4)$$

4.2 Estimation de $V^q E_{PI}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ sous l'approche MIN

Pour commencer, notons

$$E_{PI}(\hat{R}_I) \approx \frac{X^a}{N^a} \frac{X^b}{N^b},$$

où $(X^a, N^a) = \sum_{i \in U} a_i (y_i, 1)$ et $(X^b, N^b) = \sum_{i \in U} b_i (x_i, 1)$.

Par un développement en série de Taylor de premier ordre, nous pouvons montrer que $V^q E_{PI}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ peut être

approximé par

$$V^q E_{PI}(\hat{R}_I - R | \mathbf{a}, \mathbf{b}) \approx \frac{1}{N} \left[\frac{1}{1 - p^x} \left(S_y^2 + R^2 \left(1 - \frac{p^x}{p^y} \right) S_z^2 \right) - 2R \left(\frac{p^{xy}}{p^x} \frac{p^x}{p^y} \right) S_{xy}^2 \right], \quad (4.5)$$

$$S_y^2 = \frac{N}{1} \sum_{i \in U} (y_i - \bar{y})^2,$$

$$S_z^2 = \frac{N}{1} \sum_{i \in U} (x_i - \bar{x})^2$$

$$S_{xy} = \frac{N}{1} \sum_{i \in U} (x_i - \bar{x})(y_i - \bar{y})$$

$$(\bar{y}, \bar{x}) = \frac{N}{1} \sum_{i \in U} (y_i, x_i)$$

Nous obtenons un estimateur de $V^q E_{PI}(\hat{R}_I - R | \mathbf{a}, \mathbf{b})$ en estimant les quantités inconnues dans (4.5), ce qui donne

$$I_2^{(NM)} = \frac{1}{N} \left[\frac{1}{1 - \hat{p}^x} \left(S_y^2 + R^2 \left(1 - \frac{\hat{p}^x}{\hat{p}^y} \right) S_z^2 \right) - 2R \left(\frac{\hat{p}^{xy}}{\hat{p}^x} \frac{\hat{p}^x}{\hat{p}^y} \right) S_{xy}^2 \right]$$

$$= \frac{1}{N} \left[\frac{1}{1 - \frac{N^a}{N^b}} \left(S_y^2 + R^2 \left(1 - \frac{N^a}{N^b} \right) S_z^2 \right) - 2R \left(\frac{N^a}{N^b} \frac{N^b}{N^a} \right) S_{xy}^2 \right]$$

$$- \quad (4.6)$$

$$\frac{N^a}{N^b}, \hat{p}^x = \frac{N^a}{N^b} \text{ et } \hat{p}^{xy} = \frac{N^a}{N^b}.$$

L'estimateur (4.6) est asymptotiquement pI - sans biais pour la variance approximative (4.5), en notant que S_y^2 , S_z^2 et S_{xy}^2 sont asymptotiquement pI - sans biais pour S_y^2 , S_z^2 et S_{xy}^2 , respectivement. Dans le cas de l'échantillonnage aléatoire simple sans remise, l'estimateur (4.6) se réduit à

$$I_2^{(NM)} = \frac{1}{N} \left[\frac{1}{1 - \frac{N^a}{N^b}} \left(S_y^2 + R^2 \left(1 - \frac{N^a}{N^b} \right) S_z^2 \right) - 2R \left(\frac{N^a}{N^b} \frac{N^b}{N^a} \right) S_{xy}^2 \right]$$

$$- 2R \left(\frac{N^a}{N^b} \frac{N^b}{N^a} \right) S_{xy}^2 \quad (4.7)$$

Grâce à un développement en série de Taylor de premier ordre, nous obtenons

$$E^m[(E^l_r(\tilde{R}_l | s, s^{(o)}_l), s^{(p)}_l) - \tilde{R}(\tilde{R}_l | s, s^{(o)}_l), s^{(p)}_l],$$

Il en découle que V_I dans (3.6) est asymptotiquement mqI - sans biais pour V_I . Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.6) se réduit à

Un estimateur I' de $I' = E^{bdw}(I'(R) | s, s_{(x)}^d, s_{(x)}^s)$ peut s'obtenir en estimant simplement $(R) | s, s_{(x)}^d, s_{(x)}^s)$ donné par (2.6). Un estimateur asymptotiquement sans biais de $I - R | s, s_{(x)}^d, s_{(x)}^s)$ est alors donné par

$$I_{\text{NR}} = \left[\frac{1}{I} \frac{x_I}{z_I} - \frac{u}{r} \frac{r}{s} \frac{s}{z} \left(1 - \frac{u}{r} \frac{r}{s} \frac{s}{z} \right) \right] + \mathcal{H}_I^2 + \left(\frac{u}{r} \frac{r}{s} \frac{s}{z} - 1 \right) \frac{r}{s} \frac{s}{z} - \left(1 - \frac{u}{r} \frac{r}{s} \frac{s}{z} \right) \left(\frac{r}{s} \frac{s}{z} \frac{u}{x} \frac{x}{y} \frac{y}{s} \frac{s}{x} \right) \left(\frac{r}{s} \frac{s}{z} \frac{u}{x} \frac{x}{y} \frac{y}{s} \frac{s}{x} \right) \left(\frac{r}{s} \frac{s}{z} \frac{u}{x} \frac{x}{y} \frac{y}{s} \frac{s}{x} \right) \right]$$

respectivement.

Sous l'approche MI, la variance totale de \hat{R}_I peut être approximée par

$$V(\hat{R}_I - R) \approx E^q V^p E^q (R_I - R | \mathbf{a}, \mathbf{b})$$
variance totale de \hat{R}_i peut être approximée par
$$V_{(P)}^{TOT} = V_{*}^{ORD} + V^{NR} + V_I + 2V^{MIX}.$$

L'estimateur (3.8) est asymptotiquement *bdm* - sans biais pour V^{MIX} . Dans le cas de l'échantillonnage aléatoire simple sans remise, la composante V^{MIX} est nulle. De façon plus générale, la composante V^{MIX} est nulle pour tout plan de sondage autopondéré à une étape (c'est-à-dire un plan d'échantillonnage pour lequel tous les poids d'échantillonnage sont égaux). Dans le cas de plans d'échantillonnage avec probabilités inégales, il est important d'inclure la composante V^{MIX} parce que sa contribution (positive ou négative) à la variance globale peut être

$$= \mathcal{A}^{\text{XIV}} \frac{1}{z} \left\{ S_2^{\prime} \left[\frac{N_2}{\sum_{i \in s} w_i^2} - \frac{q N_2 N_2^q}{\sum_{i \in s} q_i w_i^2} \right] + S_2^{\prime} \left[\frac{N_2}{\sum_{i \in s} w_i^2} - \frac{p N_2 N_2^p}{\sum_{i \in s} p_i w_i^2} \right] \right\}$$

préservé pas la covariance, s_{xy}^{nr} dans (2.3). En effet, l'imputation a sous-estimé la relation entre les variables qui sont positivement corrélées. Donc, V^{ORD} surestime V^{SAM} , à cause de la présence du signe moins devant s_{xy}^{nr} dans (2.3). Pour surmonter cette difficulté, Samdal (1992) a proposé d'estimer $V^{DIF} = E^{nm}(V^{ORD} | s, s_{(y)}^{nr}, s_{(x)}^{nr})$ au moyen d'un estimateur mI - sans biais, V^{DIF} ; c'est-à-dire $E^{nm}(V^{DIF} | s, s_{(y)}^{nr}, s_{(x)}^{nr}) = V^{DIF}$. Toutefois, le calcul de cette composante pour un plan arbitraire comporte des opérations algébriques fastidieuses dans le cas d'un ratio. Par conséquent, nous proposons une solution de rechange qui ne nécessite aucun calcul, mais comprend la construction d'un nouvel ensemble de valeurs imputées. Elle peut être décrite comme suit : chaque fois que $a_j = 0$ et (ou) $b_j = 0$, on choisit un donneur j aléatoirement avec remise à partir de l'ensemble de répondants aux deux variables y et x (c'est-à-dire l'ensemble d'unités échantillonnées pour lesquelles $a_j = 1$ et $b_j = 1$ avec la probabilité $w_j / \sum_{i \in s} w_i a_i b_i$ et on impute le vecteur (x_j, y_j) . Autrement dit, chaque fois qu'une réponse manque pour une variable, la valeur observée est écartée et dite manquante; les valeurs manquantes sont ensuite remplacées par les valeurs d'un donneur sélectionné au hasard parmi l'ensemble de répondants aux deux variables x et y (souvent appelé ensemble de donneurs communs). De même, lorsque les valeurs des deux variables manquantes le vecteur (x_j, y_j) d'un donneur j est imputé. Puis, on applique l'estimateur de variance standard (2.2) valide sous réponse complète en utilisant ces valeurs imputées. Soit V^{ORD} l'estimateur de variance résultant. Notons que ce nouvel ensemble de valeurs imputées est utilisé pour obtenir un estimateur valide de la variance d'échantillonnage, mais ne l'est pas pour estimer le paramètre d'intérêt R . On peut montrer que V^{ORD} est un estimateur de V^{SAM} asymptotiquement $mpdI$ - sans biais. En pratique, on pourrait, par exemple, créer un fichier d'estimation de la variance contenant le nouvel ensemble de valeurs imputées et utiliser l'un des systèmes standard d'estimation de la variance (employés dans le cas de données complètes) pour obtenir l'estimation de la variance d'échantillonnage.

3.2 Estimation de la variance de non-réponse V^{NR}

Un estimateur V^{NR} de $V^{NR} = E^{pq}(V^{nm}(E(R_j | s, s_{(y)}^{nr}, s_{(x)}^{nr}) - R | s, s_{(y)}^{nr}, s_{(x)}^{nr})) - R | s, s_{(y)}^{nr}, s_{(x)}^{nr})$. Par un développement en série de Taylor de premier ordre, nous arrivons à

$$V^{NR} \approx \left\{ \frac{1}{\sum w_i^2 a_i} \frac{N^2}{\sum w_i^2 a_i} + \frac{N^2}{\sum w_i^2 a_i} - 2 \frac{N^2}{\sum w_i^2 a_i} \frac{N^2}{\sum w_i^2 a_i} \right\} \sigma_x^2 + \left\{ \frac{1}{\sum w_i^2 b_i} \frac{N^2}{\sum w_i^2 b_i} + \frac{N^2}{\sum w_i^2 b_i} - 2 \frac{N^2}{\sum w_i^2 b_i} \frac{N^2}{\sum w_i^2 b_i} \right\} \sigma_y^2 + 2 \left\{ \frac{1}{\sum w_i^2 a_i b_i} \frac{N^2}{\sum w_i^2 a_i b_i} + \frac{N^2}{\sum w_i^2 a_i b_i} - 2 \frac{N^2}{\sum w_i^2 a_i b_i} \frac{N^2}{\sum w_i^2 a_i b_i} \right\} \sigma_{xy}^2 \quad (3.4)$$

où $(N_j, N_a, N_b) = \sum_{i \in s} w_i (1, a_i, b_i)$. Maintenant, posons que $s_x^2 = 1 / N \sum_{i \in s} w_i (x_i - \bar{x})^2$ et $s_y^2 = 1 / N \sum_{i \in s} w_i (y_i - \bar{y})^2$ avec $(\bar{x}, \bar{y}) = (1 / N \sum_{i \in s} w_i x_i, 1 / N \sum_{i \in s} w_i y_i)$. Notons que s_x^2 et s_y^2 désignent respectivement la variabilité d'échantillon et s_{xy}^2 la covariance sous le modèle σ_{xy}^{nr} . Il s'ensuit que V^{NR} s'obtient par estimation des quantités inconnues dans (3.4), ce qui mène à

L'estimateur (3.5) est asymptotiquement $mpdI$ - sans biais pour V^{NR} . Dans le cas particulier de l'échantillonnage aléatoire simple sans remise, l'expression (3.5) se réduit à

$$V^{NR} = \left\{ \frac{1}{\sum w_i^2 a_i} \frac{N^2}{\sum w_i^2 a_i} + \frac{N^2}{\sum w_i^2 a_i} - 2 \frac{N^2}{\sum w_i^2 a_i} \frac{N^2}{\sum w_i^2 a_i} \right\} \sigma_x^2 + \left\{ \frac{1}{\sum w_i^2 b_i} \frac{N^2}{\sum w_i^2 b_i} + \frac{N^2}{\sum w_i^2 b_i} - 2 \frac{N^2}{\sum w_i^2 b_i} \frac{N^2}{\sum w_i^2 b_i} \right\} \sigma_y^2 + 2 \left\{ \frac{1}{\sum w_i^2 a_i b_i} \frac{N^2}{\sum w_i^2 a_i b_i} + \frac{N^2}{\sum w_i^2 a_i b_i} - 2 \frac{N^2}{\sum w_i^2 a_i b_i} \frac{N^2}{\sum w_i^2 a_i b_i} \right\} \sigma_{xy}^2 \quad (3.5)$$

3. Estimation de la variance : cadre à deux phases

À la présente section, nous obtenons des estimateurs de la variance sous le cadre à deux phases et l'approche MI, suivant la méthode proposée par Särndal (1992), ainsi que par Deville et Särndal (1994). En utilisant la décomposition (2.8), la variance totale de R_I peut être approximée par

$$V^{mpqI}(R_I - R) \approx E^{mpqI}(R_I - R)^2$$

$$= V^{SAM} + V^{NR} + V^{I} + 2V^{NIX}, \quad (3.1)$$

où $V^{SAM} = E^{SM}(\hat{R}) = E^{SM}(V^{SAM})$ est la variance d'échantillonnage de l'estimateur sous données complètes R , $V^{NR} = E^{pd} V^m(E_I(R_I | s, s'_{(v)}, s'_{(x)})) - R | s, s'_{(v)}, s'_{(x)})$ est la variance de non-réponse de l'estimateur imputé R_I , $V^I = E^{mpq} V^I(R_I - E_I(R_I | s, s'_{(v)}, s'_{(x)})) | s, s'_{(v)}, s'_{(x)})$ est la variance due à l'imputation de l'estimateur imputé R_I , et $V^{NIX} = E^{pqmI}[(E_I(R_I | s, s'_{(v)}, s'_{(x)})) - R)(R - R) | s, s'_{(v)}, s'_{(x)}]$ une composante mixte. Notons que l'expression (3.1) ne contient qu'un seul terme de produit croisé, $2V^{NIX}$, parce que tous les autres sont asymptotiquement nuls.

3.1 Estimation de la variance d'échantillonnage

Soit V^{ORD}_{SAM} l'estimateur de variance naïf de R_I , c'est-à-dire l'estimateur de variance obtenu en traitant les valeurs imputées comme s'il s'agissait de valeurs observées. Nous parvenons à cet estimateur en remplaçant e_i par $\hat{e}_i = 1/X_I(Y_i - R_I X_I)$ dans (2.2), ce qui donne

$$V^{ORD} = \sum_{i \in s} \sum_{j \in s} \Delta_{ij} \hat{e}_i \hat{e}_j. \quad (3.2)$$

Comme nous le montrons maintenant dans le cas de l'échantillonnage simple sans remise, V^{ORD}_{SAM} sous l'imputation HDAM quand $\sigma_{en} > 0$ (comme cela est habituellement le cas en pratique). Après certains calculs algébriques, nous obtenons

$$E^{mm}(V^{ORD} - V^{SAM} | s, s'_{(v)}, s'_{(x)}) \approx \frac{1}{2} \left(\frac{H_x}{H_y} \right) \left(1 - \frac{N}{n} \right) \left(1 - \frac{u}{r_{xy}} \right) \left(\frac{u}{\sigma_{en}} \right). \quad (3.3)$$

L'expression (3.3) montre que V^{ORD} est $mpqI$ - biaisé pour V^{SAM} , à moins que $\sigma_{en} = 0$, $r_{xy} = n$ (ce qui est le cas des données complètes) ou que $n = N$ (ce qui est le cas d'un recensement). Le fait que V^{ORD} ne soit pas un estimateur valide de V^{SAM} s'explique facilement en notant que, si l'imputation HDAM préserve la variabilité, c'est-à-dire s^2_x et s^2_y , correspondant aux variables x et y , elle ne

$$m: \begin{cases} Y_i = \mu_y + \varepsilon_i \\ X_i = \mu_x + \eta_i \end{cases} \quad (2.7)$$

où ε_i est un terme d'erreur aléatoire tel que $E^m(\varepsilon_i) = 0$, $E^m(\varepsilon_i \varepsilon_j) = 0$, pour $i \neq j$, $V^m(\varepsilon_i) = \sigma^2_\varepsilon$ et η_i est un terme d'erreur aléatoire tel que $E^m(\eta_i) = 0$, $E^m(\eta_i \eta_j) = 0$, pour $i \neq j$, $V^m(\eta_i) = \sigma^2_\eta$. En outre, nous supposons que $\{w_i; i \in U\}$ ni $\{\mathbf{II}^U = \{w_i; i \in U\}$ ne dépend pas de $s, s'_{(v)}, s'_{(x)}$, $\mathbf{W}^U = \{\mathbf{II}^U; i \in U\}$ que la distribution des erreurs $(\varepsilon_i, \eta_i); i \in U\}$ est conditionnelle à $\{w_i; i \in U\}$, après conditionnement sur \mathbf{Z}^U . Par conséquent, sauf les variables d'intérêt y et x , toutes les variables qui interviennent dans l'estimateur imputé (2.4) sont traitées comme étant fixes lorsqu'on considère les espérances et les variances par rapport au modèle d'imputation.

2.3 Biais de l'estimateur imputé

Pour étudier le biais de l'estimateur imputé (2.4), nous utilisons la décomposition standard de l'erreur totale de

$$R_I - R = [R - R] + [E_I(R_I | s, s'_{(v)}, s'_{(x)}) - R] + [R_I - E_I(R_I | s, s'_{(v)}, s'_{(x)})]. \quad (2.8)$$

Le premier terme $R - R$ du deuxième membre de (2.8) est appelé erreur d'échantillonnage, le deuxième terme $E_I(R_I | s, s'_{(v)}, s'_{(x)}) - R$ est appelé erreur de non-réponse, tandis que le troisième terme $R_I - E_I(R_I | s, s'_{(v)}, s'_{(x)})$ est appelé erreur d'imputation.

À l'aide d'un développement en série de Taylor du premier ordre, il est facile de montrer que, sous l'approche MN, l'estimateur imputé (2.4) est asymptotiquement pqI - sans biais, c'est-à-dire que $E^{pqI}(R_I - R) \approx 0$. De plus, sous l'approche MI et le modèle (2.7), nous pouvons montrer que l'estimateur imputé (2.4) est asymptotiquement $mpqI$ - sans biais, c'est-à-dire que $E^{mpqI}(R_I - R) \approx 0$. Donc, l'estimateur imputé est robuste en ce sens qu'il est valide sous l'approche MN ou sous l'approche MI. Notons que, pour que le biais asymptotique soit nul sous les deux approches, nous imposons que la taille d'échantillon soit suffisamment grande dans chaque classe d'imputation. Dans ce qui suit, nous supposons donc que le biais de R_I est négligeable.

$$E_I(\hat{R}_I | s, s_{(y)}^*, s_{(x)}^*) \approx \frac{\hat{X}_I}{\hat{Y}_I} \equiv \hat{R}_I, \quad (2.5)$$

$$\frac{\hat{Y}_I}{\hat{X}_I} = \frac{\sum_{i \in s} w_i a_i y_i}{\sum_{i \in s} w_i b_i y_i}$$

où

et

$$\frac{\hat{X}_I}{\hat{Y}_I} = \frac{\sum_{i \in s} w_i b_i y_i}{\sum_{i \in s} w_i b_i}$$

désignent les moyennes pondérées des répondants pour les variables y et x , respectivement. L'approximation (2.5) sera valide si la taille d'échantillon dans les classes est suffisamment grande, ce que nous supposons être le cas.

Maintenant, soient

$$s_y^* = \frac{1}{\sum_{i \in s} w_i a_i} \sum_{i \in s} w_i a_i (y_i - \bar{y}_I)$$

et

$$s_x^* = \frac{1}{\sum_{i \in s} w_i b_i} \sum_{i \in s} w_i b_i (x_i - \bar{x}_I)$$

la variabilité des valeurs de y et des valeurs de x dans les ensembles de répondants $s_{(y)}^*$ et $s_{(x)}^*$, respectivement. En notant que, sous l'imputation HDAM pondérée,

$$I_I(y_i^*) = s_y^* = s_x^*$$

et

$$\text{COV}_I(y_i^*, x_i^*) = 0,$$

nous pouvons approximer $I_I(\hat{R}_I | s, s_{(y)}^*, s_{(x)}^*)$ par

$$I_I(\hat{R}_I | s, s_{(y)}^*, s_{(x)}^*)$$

$$\approx \frac{1}{\sum_{i \in s} w_i^2} \left[\sum_{i \in s} w_i^2 (1 - a_i) s_y^* + \hat{R}_I^2 \sum_{i \in s} w_i^2 (1 - b_i) s_x^* \right]. \quad (2.6)$$

Les expressions (2.5) et (2.6) seront utilisées aux sections suivantes lors de la discussion du biais et de la variance de l'estimateur imputé \hat{R}_I . Comme nous le verrons aux sections 3 et 4, la variance conditionnelle (2.6) est une mesure de la variabilité due au mécanisme d'imputation. Nous allons maintenant décrire deux approches d'intervariance aux sections 3 et 4, à savoir l'approche du modèle de non-réponse (MN) et l'approche du modèle d'imputation (MI).

2.2 Approche du modèle d'imputation

Dans l'approche MI, l'inférence est faite par rapport à la distribution conjointe induite par le modèle d'imputation, le plan d'échantillonnage et le modèle de non-réponse. Le modèle d'imputation est un ensemble d'hypothèses au sujet de la loi inconnue de (Y_i, X_i) , $i \in U$. Dans une classe d'imputation, sous imputation HDAM, le modèle d'imputation, m , est donné par

Dans l'approche MN, l'inférence est faite par rapport à la distribution conjointe induite par le plan d'échantillonnage et le modèle de non-réponse. Ce dernier est un ensemble d'hypothèses au sujet de la loi inconnue des indicateurs de réponse $\mathbf{R}_s = \{(a_i, b_i); i \in s\}$. Cette loi inconnue est souvent appelée mécanisme de non-réponse. Soit $P^{y_I} = P(a_i = 1 | s, \mathbf{Z}_s)$, la probabilité de réponse de l'unité i pour la variable y , où $\mathbf{Z}_s = \{\mathbf{z}_i; i \in s\}$ et \mathbf{z}_i est un vecteur de variables auxiliaires disponibles pour toutes les unités de l'échantillon utilisées pour former les classes d'imputation. De même, soit $P^{x_I} = P(b_i = 1 | s, \mathbf{Z}_s)$ la probabilité de réponse de l'unité i pour la variable x . Nous supposons que les unités répondent indépendamment les unes des autres; c'est-à-dire que $P^{y_I} = P(a_i = 1, b_j = 1 | s, \mathbf{Z}_s) = P^{y_I} P^{x_I}$ pour $i \neq j$ et $P^{x_I} = P(b_i = 1, b_j = 1 | s, \mathbf{Z}_s) = P^{x_I} P^{y_I}$ pour $i \neq j$. Cependant, nous ne supposons pas que, pour une unité donnée i , la réponse pour la variable y est indépendante de celle pour la variable x . Autrement dit, si nous posons que $P^{y_I} = P(a_i = 1 | s, \mathbf{Z}_s)$, nous avons alors $P^{x_I} \neq P^{y_I} P^{x_I}$ en général. Dans une classe d'imputation, nous supposons que le mécanisme de réponse est uniforme, de sorte que $P^{y_I} = P^{x_I} P^{x_I}$ et $P^{x_I} = P^{y_I} P^{x_I}$.

2.1 Approche du modèle de non-réponse

Estimation de la variance pour un ratio en présence de données imputées

David Haziza

Résumé

Dans le présent article, nous étudions le problème de l'estimation de la variance pour un ratio de deux totaux quand l'imputation hot deck aléatoire marginale est utilisée pour remplacer les données manquantes. Nous considérons deux approches d'inférence. Dans la première, l'établissement de la validité d'un modèle d'imputation est nécessaire. Dans la seconde, la validité d'un modèle d'imputation n'est pas nécessaire, mais il faut estimer les probabilités de réponse, auquel cas il est nécessaire d'établir la validité d'un modèle de non-réponse. Nous obtenons les estimateurs de la variance sous deux cadres distincts, à savoir le cadre à deux phases habituel et le cadre inverse.

Mots clés : Modèle d'imputation; modèle de non-réponse; imputation hot deck aléatoire marginale; cadre inverse; cadre à deux phases; estimation de la variance.

1. Introduction

L'estimation de la variance en présence de données imputées pour des paramètres univariés simples, tels que des totaux ou des moyennes de population, a fait l'objet de nombreuses études ces dernières années; voir, par exemple, Samdal (1992), Deville et Samdal (1994), Rao et Shao (1992), Rao (1996), et Shao et Steel (1999). Il est fréquent, en pratique, de devoir estimer le ratio de deux totaux de population, $R = Y/X$, où $(Y, X) = \sum_{i \in U} (y_i, x_i)$, y et x sont deux variables d'intérêt pour lesquelles des données peuvent manquer et U est la population finie (de taille N) étudiée. Alors que l'estimation de la variance dans le cas d'un ratio en présence de données imputées est un problème qui se pose souvent en pratique (surtout dans le cas des enquêtes auprès des entreprises), autant que nous sachions, la question n'a pas été étudiée à fond dans la littérature. Dans le présent article, nous considérons le cas de l'imputation hot deck aléatoire marginale (HDAM) effectuée dans le même ensemble de classes d'imputation pour les deux variables y et x . Autrement dit, pour corriger la non-réponse, on procède séparément à une imputation hot deck aléatoire pour chaque variable dans le même ensemble de classes d'imputation. Cette situation se présente fréquemment en pratique. Pour simplifier, nous considérons le cas d'une seule classe d'imputation. Les extensions aux classes d'imputation multiples sont relativement simples pour la plupart des calculs présentés ici.

Nous obtenons dans le présent article des estimateurs de la variance qui tiennent compte de l'échantillonnage, de la non-réponse et de l'imputation. Deux cadres distincts d'estimation de la variance ont été étudiés dans la littérature: i) le cadre à deux phases habituel (par exemple, Samdal (1992)) et ii) le cadre inversé (par exemple, Shao et Steel (1999)). Dans le cadre à deux phases, la non-réponse est considérée

comme une deuxième phase de sélection. En d'autres termes, un échantillon aléatoire est sélectionné parmi la population selon le plan d'échantillonnage donné. Puis, étant donné l'échantillon sélectionné, l'ensemble de réponses est généré selon le mécanisme de non-réponse. Dans le cadre inverse, l'ordre de l'échantillonnage et de la réponse est inverse. Autrement dit, la population est d'abord répartie aléatoirement en une population de répondants et une population de non-répondants, selon le mécanisme de non-réponse. Puis, un échantillon aléatoire est sélectionné à partir de la population (contenant les répondants et les non-répondants), selon le plan d'échantillonnage. Comme nous le verrons à la section 4, le cadre inverse facilite le calcul des estimations de la variance, mais contrairement au cadre à deux phases, nécessite l'hypothèse supplémentaire que le mécanisme de non-réponse ne dépend pas de l'échantillon sélectionné. Cette hypothèse est satisfaisante dans de nombreuses situations observées en pratique. Pour chaque cadre, l'inférence peut être fondée sur un modèle d'imputation (MI) ou sur un modèle de non-réponse (MN). L'approche MI requiert la validité d'un modèle d'imputation et l'approche MN requiert la validité d'un modèle de non-réponse.

Skinner, C.J. (1991). On the efficiency of raking ratio estimator for multiple frame surveys. *Journal of American Statistical Association*, 86, 779-784.

Skinner, C.J., Holmes, D.J. et Holt, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 333-347.

Skinner, C.J., et Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of American Statistical Association*, 91, 349-356.

Sudman, S., et Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.

Sudman, S., Sirken, M.G. et Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 2, 991-996.

Bibliographie

Bankier, M.D. (1986). Estimators based in several stratified samples with applications to multiple frame surveys. *Journal of American Statistical Association*, 81, 1074-1079.

Casady, R.J., et Siken, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 601-605.

Carlson, B.L., et Hall, J.W. (1994). Weighting sample data when multiple sample frames are used. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 882-887.

Deville, J.C., et Sîndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of American Statistical Association*, 87, 376-382.

Deville, J.-C., Sîndal, C.-E., et Satoru, O. (1993). Generalized raking procedures in survey sampling. *Journal of American Statistical Association*, 88, 1013-1020.

Eurostat (2000). *Push and Pull Factors of International Migration*. Country Report-Italy, 3/2000/E/n.5, Bruxelles: European Communities Printing Office.

Fuller, W.A., et Burnmeister, L.F. (1972). Estimators of samples selected from two overlapping frames. *Proceedings of the Social Statistics Sections*, American Statistical Association, 245-249.

Haines, D.E., et Pollock, K.H. (1998). Combinaison de bases multiples pour estimer la taille et les chiffres de la population. *Techniques d'enquête*, 24, 81-91.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Sections*, American Statistical Association, 203-206.

Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Samkhyā*, 36, 99-118.

Iachan, R., et Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.

Kalton, G., et Anderson, D.W. (1986). Sampling rare populations. *Journal of Royal Statistical Society, A*, 149, 65-82.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles, Éditions Ellipses.

Lavallée, P. (2007). *Indirect Sampling*. New York : Springer.

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of American Statistical Association*, 95, 271-280.

Lohr, S., et Rao, J.N.K. (2006). Multiple frame surveys: Point estimation and inference. *Journal of American Statistical Association*, 101, 1019-1030.

Lund, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Sections*, American Statistical Association, 282-288.

Mecatti, F. (2004). Échantillonnage de centres : stratégie d'enquête auprès des populations difficiles à échantillonner. *Recueil du Symposium de Statistique Canada*.

Mecatti, F. (2005). Estimation d'après une base de sondage unique dans une enquête à base de sondage multiple. *Recueil du Symposium de Statistique Canada*.

Siken, M.G. (2004). Enquêtes auprès de populations rares ou difficiles à rejoindre au moyen d'échantillonnage par réseau : revue historique. *Recueil du Symposium de Statistique Canada*, Discours principal.

comparativement à \hat{Y}_M pour la classification exacte, c'est-à-dire le cas d'un chevauchement/d'une couverture élevée des bases de sondage et d'une faible disproportion d'échantillonnage. Dans ces conditions, nous avons étudié des taux croissants d'erreur de classification des unités échantillonnées dans les domaines (variant de 0 à 50 %). Les tableaux 6 et 7 montrent, respectivement, le biais relatif et le ratio d'efficacité de \hat{Y}_M par rapport aux estimateurs BU, BUMiq et PMV, pour des niveaux croissants d'erreur de classification. Il convient de souligner que, si les effets négatifs de l'erreur de classification sont rapides et importants pour tous les estimateurs étudiés, l'estimateur PMV est celui qui est le moins affecté.

En conclusion, outre le fait qu'il est simple, l'estimateur fondé sur la multiplicité proposé est recommandé lorsque le risque (même léger) d'erreur de classification des unités échantillonnées dans les domaines est une réelle possibilité.

Tableau 6
Biais relatif (absolu) pour un taux croissant d'erreur de classification

Erreur de classification en %		\hat{Y}_M	BU	BUMiq	PMV
0	0	0	0	0	0
1 %	0	2,57	1,38	4,3	0
5 %	0	13,57	7,15	2,75	0
10 %	0	17,80	14,14	4,56	0
20 %	0	25	25	6	0
50 %	0	144	68	39	0

Tableau 7
Ratio d'efficacité empirique de \hat{Y}_M vs. les estimateurs BU, BUMiq et PMV pour un taux croissant d'erreur de classification

Erreur de classification en %		\hat{Y}_M vs. BU	\hat{Y}_M vs. BUMiq	\hat{Y}_M vs. PMV
0	0,640	3,210	3,300	1,100
1 %	0,260	1,400	0,370	0,150
5 %	0,020	0,060	0,020	0,010
10 %	0,010	0,020	0,004	0,004
20 %	0,004	0,004	0,001	0,001
50 %	≈ 0	0,006	0,006	0,006

Remerciements

Les présents travaux ont été financés en partie par une subvention du ministère italien des Universités et de la Recherche. L'auteur tient à remercier Jon N.K. Rao pour sa discussion fructueuse et ses conseils utiles. Elle remercie également le rédacteur en chef, deux rédacteurs adjoints et deux examinateurs anonymes de leurs commentaires et suggestions constructives.

MIQ) à des niveaux croissants d'erreur de classification des unités échantillonnées dans les domaines, comparativement à l'insensibilité structurelle de l'estimateur fondé sur la multiplicité proposé. Pour un taux donné d'erreur de classification, le nombre souhaité d'unités échantillonnées qui doivent être classées incorrectement est tiré du domaine dont la taille est la plus grande et assigné aléatoirement aux autres domaines, indépendamment pour chaque base de

Les tableaux 4 et 5 donnent les statistiques élémentaires résumant les résultats des simulations dans le cas d'une classification exacte et dans celui d'une légère erreur de classification égale à 1 % des unités échantillonnées. Notons que, pour la classification exacte, tous les estimateurs paraissent être sans biais (ou presque sans biais). En ce qui concerne l'efficacité, selon d'autres résultats de simulation (Lohr et Rao 2006), les estimateurs BUmig et PMV ont des propriétés comparables. Comme prévu, pour la classification exacte, ils sont plus efficaces que \hat{Y}_M dans tous les cas étudiés (sauf deux cas isolés), à cause de la quantité différente d'information utilisée dans le processus d'estimation. Cependant, les estimateurs BU (simple et ajusté par la MIQ) et PMV ont tendance à devenir biaisés et moins efficaces que \hat{Y}_M en présence d'un faible taux d'erreur de classification.

Tableau 4
Biais relatif pour une erreur de classification de 1 % : statistiques élémentaires sur 29 populations simulées

BR _{mc} (absolu)	Moyenne	Min.	Max.	Médiane	75 ^e quantile
d'unités échantillonnées pour 1 %					
\hat{Y}_M	0	0	0	0	0
BU	2,5880	0,83	7,02	2,65	2,13
BUmig	1,7632	0,73	4	1,97	1,65
PMV	2,7352	0,23	4,67	3,46	2,87

Tableau 5
Ratio d'efficacité empirique de \hat{Y}_M vs. les estimateurs BU et PMV : statistiques élémentaires sur les 29 populations simulées pour la classification exacte et pour une légère erreur de classification

Ratio d'efficacité empirique	Moyenne	Min.	Max.	Médiane	75 ^e quantile
Classification exacte					
BUmig	1,43	0,69	3,21	1,51	1,28
PMV	1,41	0,72	3,30	1,47	1,25
Erreur de classification de 1 %					
BU	0,39	0,13	0,71	0,54	0,34
BUmig	0,78	0,13	1,98	0,95	0,74
PMV	0,77	0,14	1,94	0,98	0,70

Enfin, nous nous concentrons sur le cas de l'efficacité maximale des estimateurs BU, BUmig et PMV

Nous avons étudié différents niveaux de couverture des fractions d'échantillonnage dominant lieu à une disproportion croissante d'échantillonnage, dans les simulations.

Les résultats empiriques montrent que Mmig est plus efficace que BUmig dans 38 % des cas étudiés et qu'il est aussi ou moins efficace dans les autres cas. Les gains d'efficacité varient de 3 % à 74 % et ont lieu pour les faibles niveaux de couverture des bases de sondage. Lorsque la couverture de la base de sondage augmente (et donc également le chevauchement), l'estimateur Mmig est supérieur à l'estimateur BUmig en cas de forte disproportion d'échantillonnage seulement. Dans les autres conditions, c'est-à-dire un accroissement de la couverture et du chevauchement des bases de sondage, combiné à une disproportion faible à moyenne d'échantillonnage, Mmig peut être sensiblement moins efficace que BUmig (voir le tableau 3 pour dix cas indicatifs) et fortement biaisé. Les résultats empiriques laissent donc entendre que l'ajustement par la MIQ a de meilleurs effets sous une approche de base de sondage unique simple que sous une approche de multiplicité, bien qu'il existe des conditions dans lesquelles cette dernière demeure supérieure. Il faudra poursuivre les travaux de recherche à cet égard. En particulier, puisque la MIQ est en fait un cas particulier de calage (Deville et Särndal 1992; Deville, Särndal et Sautory 1993), les résultats pourraient être améliorés en appliquant la méthode plus générale de calage à l'estimateur \hat{Y}_M . Le calage de l'estimateur fondé sur la multiplicité, considéré comme un cas particulier de la méthode généralisée de partage des poids, est décrit dans Lavallée (2002, 2007).

Tableau 3
Efficacité de Mmig vs. BUmig : dix exécutions de simulation indicatives

Couverture de la base de sondage	$\alpha_g = n_g/N$	Fraction d'échantillonnage $f_g = n_g/N_g$	Ratio d'efficacité empirique Mmig vs. BUmig
0,60	0,60	0,01	0,26
0,35	0,35	0,80	0,54
0,85	0,85	0,01	0,71
0,35	0,40	0,70	0,96
0,85	0,85	0,80	1,01
0,60	0,60	0,70	1,09
0,80	0,50	0,01	1,22
0,35	0,40	0,95	1,63
0,70	0,05	0,70	2,09
0,70	0,05	0,80	5,79

4.4 Erreur de classification

Le but de la dernière partie de l'étude par simulation est d'étudier la sensibilité des estimateurs pseudo-optimal (PMV) et à base de sondage unique (simple et ajusté par la

Tableau 1 Ratio d'efficacité empirique de Y_M vs. l'estimateur BL : statistiques élémentaires sur 28 populations simulées				
Moyenne	Max.	Min.	Médiane	75 ^e quantile
0,7425	0,95	0,52	0,74	0,89

basées de sondage.

Tableau 2 Ratio d'efficacité empirique de Y_M vs. l'estimateur BU pour une disproportion d'échantillonnage croissante	
Disproportion d'échantillonnage	Ratio d'efficacité empirique moyen
0,11 0,22 0,31 0,40	0,92 0,81 0,68 0,57
des bases de sondage	

4.3 Ajustement par la méthode itérative du quotient (MIQ)

D'aucuns ont proposé de procéder à un ajustement par la MIQ en utilisant les tailles connues des bases de sondage N_q (Bankier 1986) afin d'accroître l'efficacité de l'estimateur BU simple. Les résultats théoriques et empiriques déjà publiés confirment que l'estimateur BU ajusté par la MIQ (BUMiq) peut être considérablement plus efficace que l'estimateur BU simple (Skinner 1991; Lohr et Rao 2000, 2006; Mecatti 2005).

Afin de corriger l'estimateur fondé sur la multiplicité par la MIQ, nous devons émettre l'hypothèse que nous connaissons l'appartenance des unités échantillonnées aux domaines. En utilisant cette information supplémentaire, quoique redondante, nous pouvons récrire Y_M sous la forme

$$Y_M = \sum_{q \in K} \sum_{i \in q} (|K| f_q^i)^{-1} \sum_{i \in q} \delta_i(K) y_i \quad (7)$$

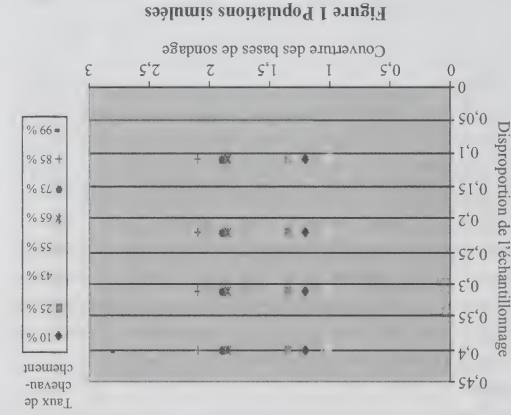
où $|K|$ indique le nombre de bases de sondage intervenant dans le domaine D_K et est égal à la multiplicité m_i de l'unité pour tout $i \in D_K$. En fixant les poids initiaux à $h_{(0)}^{Kq} = (|K| f_q^i)^{-1}$, nous obtenons la i^{e} itération de l'estimateur fondé sur la multiplicité corrigé par la MIQ (Mimiq) en introduisant les poids corrigés suivants par substitution dans (7)

$$h_{Kq}^{(i)} = \begin{cases} \frac{\sum_{q \in K} h_{Kq}^{(i-1)} n_q}{N^{b_{Kq}} h_{Kq}^{(i-1)}} & \text{si } q \in K \\ h_{Kq}^{(i-1)} & \text{si } q \notin K \end{cases}$$

autrement, pour $i = 1, 2, \dots$ jusqu'à la convergence.

dans le plan formé par les deux principaux paramètres de simulation, à savoir la couverture (totale) des bases de sondage sur l'axe horizontal (donnée par $\sum q^q \alpha_q^i$) et la disproportion de l'échantillonnage sur l'axe vertical, c'est-à-dire la dispersion des fractions d'échantillonnage f_q^i , mesurée par $\sum q^q |f_q^i - f_q^j| / 3^2$. Dans la figure 1, les divers formes des symboles représentant les populations/points indiquent divers niveaux de chevauchement, c'est-à-dire le taux total d'unités de population classées dans les quatre domaines chevauchants.

4.2 Estimation en base de sondage unique fondée sur la multiplicité comparativement à l'approche simple



Comme il est mentionné à la section 3, l'estimateur fondé sur la multiplicité fait intervenir des poids particuliers à la base de sondage, tandis que l'estimateur BU simple est fondé sur des coefficients moyens. Par conséquent, les deux estimateurs coïncident si la fraction d'échantillonnage $f_q^i = f$ est la même dans chaque base de sondage, c'est-à-dire pour un échantillonnage proportionnel, et ils produisent des estimations différentes en cas d'échantillonnage disproportionnel. Les résultats de simulation fournissent des preuves empiriques que l'estimateur fondé sur la multiplicité est plus exact que l'estimateur BU simple. Ils montrent que l'estimateur Y_M est plus efficace dans tous les cas examinés, sauf un cas extrême où les trois bases de sondage sont presque complètes et que le chevauchement total est proche de 100%. Si l'on omet ce cas particulier, le gain d'efficacité de Y_M par rapport à l'estimateur BU, mesuré par un ratio d'efficacité empirique ordinaire (voir le tableau 1), varie de 5 % à 48 %, et n'est jamais inférieur à 26 % dans la moitié des simulations. L'efficacité de l'estimateur fondé sur la multiplicité par rapport à l'estimateur BU simple augmente à mesure que s'accroît la disproportion de l'échantillonnage (voir le tableau 2), tandis

(iii) étudier les effets de l'accroissement des taux d'erreur de classification sur les propriétés empiriques des estimateurs PMV et BU (simple et ajusté par la MIO) comparativement à l'insensibilité naturelle de l'estimateur fondé sur la multiplicité (section 4.4).

4.1 Mise en œuvre

L'étude par simulation a été exécutée dans des conditions artificielles de sondage à trois bases et mise en œuvre comme il suit. N pseudo-valeurs de population y_i sont générées à partir d'une loi gamma. Certains simulations préliminaires ont indiqué que ni l'accroissement de la taille de la population N ni l'utilisation de valeurs différentes pour les paramètres gamma (produisant une forme asymétrique et une forme presque symétrique) ne font varier sensiblement le profil de performance relative des estimateurs considérés. L'étude a été réalisée en fixant $N = 1\,200$ et en générant les valeurs à partir d'une loi gamma dont les paramètres étaient 1,5 et 2. Chaque pseudo-valeur y_i est affectée aléatoirement aux $\bar{Q} = 3$ bases de sondage conformément à 3 essais de Bernoulli indépendants avec probabilité $\alpha_q = N_q/N$, $q = 1, 2, 3$. Plusieurs scénarios concernant la couverture ainsi que le chevauchement des bases de sondage ont été obtenus en choisissant diverses valeurs de α_q sous deux contraintes : a) $\sum \alpha_q \geq 1$ pour s'assurer que les trois bases couvrent l'entière de la population et b) aucune des trois bases n'est vide. Dans certains cas, le chevauchement soulnaité des bases de sondage a été produit en fixant le ratio N_q/N des unités de population incluses dans chaque domaine.

Étant donné un ensemble de fractions d'échantillonnage $f_q = n_q/N$, $q = 1, 2, 3$, un échantillon aléatoire simple est sélectionné indépendamment dans chaque base de sondage, itérativement pour 10 000 exécutions de la simulation. Pour un estimateur donné, disons \hat{Y} , il est supposé que l'ensemble de valeurs $\{\hat{Y}^p, p = 1, \dots, 10\,000\}$ correspond à sa distribution de Monte Carlo et l'on calcule la moyenne empirique $E_{mc}(\hat{Y}) = \sum \hat{Y}^p / 10\,000$ et l'erreur quadratique moyenne empirique $EQM_{mc}(\hat{Y}) = \sum (\hat{Y}^p - E_{mc}(\hat{Y}))^2 / 10\,000$. L'erreur de Monte Carlo est contrôlée en acceptant uniquement les simulations produisant un biais relatif empirique $BR_{mc}(\hat{Y}) = 100 \cdot |E_{mc}(\hat{Y}) - Y| / Y$ inférieur à 1,5 % pour les estimateurs que l'on sait être sans biais. En outre, en utilisant la variance exacte de l'estimateur fondé sur la multiplicité donnée par (6), on s'assure que les simulations donnentont $|EQM_{mc}(\hat{Y}_M) - V(\hat{Y}_M)| \leq 0,03$. Plusieurs scénarios ont été étudiés en combinant divers niveaux de couverture des bases de sondages, de chevauchement des bases de sondage et de disproportion de l'échantillonnage, ce qui aboutit à la production de 29 populations simulées. À la figure 1, ces populations sont représentées par des points

$$V(\hat{Y}_M) = \frac{\sum_{q=1}^b N_q^b - n_q^b}{\sum_{q=1}^b N_q^b} \left[N^b \sum_{i \in A_q} y_i^2 m_{i-2} - \left(\sum_{i \in A_q} y_i m_{i-1} \right)^2 \right]. \quad (5)$$

Un estimateur sans biais de la variance sous échantillonnage aléatoire simple de chaque base de sondage est alors donné par

$$\hat{V}(\hat{Y}_M) =$$

$$\frac{\sum_{q=1}^b N_q^b (N_q^b - n_q^b)}{N^b} \left[N^b \sum_{i \in A_q} y_i^2 m_{i-2} - f_q^{-1} \left(\sum_{i \in A_q} y_i m_{i-1} \right)^2 \right]. \quad (6)$$

Les propriétés de l'estimateur fondé sur la multiplicité pour des tailles finies d'échantillon ont été étudiées empiriquement pour l'échantillonnage aléatoire simple et comparées à celles des principaux estimateurs concurrents par une méthode de simulation.

4. Étude par simulation

Plusieurs résultats de simulation concernant les estimateurs pour base de sondage double ont été publiés (Bankier 1986; Skinner et Rao 1996; Lohr et Rao 2000). Dans le cas général de $\bar{Q} \geq 2$ bases de sondage, Lohr et Rao (2006) ont étudié en détail les erreurs quadratiques moyennes empiriques d'un groupe de huit estimateurs sous les approches optimale, pseudo-optimale et pour base de sondage unique, dans un cadre à trois bases de sondage sous un plan de sondage à deux degrés. Leurs résultats donnent à penser que les estimateurs optimaux le sont en théorie, mais qu'en pratique, la variabilité supplémentaire due à l'estimation des poids optimaux produit des erreurs quadratiques moyennes plus importantes. Donc, l'estimateur PMV semble être celui qui donne les meilleurs résultats en ce qui concerne l'efficacité relative empirique. En outre, l'étude comportait l'examen d'un cas où environ 10 % des unités échantillonnées avaient été classées incorrectement dans les domaines, ce qui a poussé les auteurs à recommander d'étudier plus en profondeur les effets des erreurs de classification sur les propriétés de l'estimateur. Dans notre étude, nous comparons les estimateurs pseudo-optimal et pour base unique à l'estimateur fondé sur la multiplicité (4) en poursuivant trois grands objectifs :

- i) déterminer les conditions empiriques dans lesquelles l'estimateur fondé sur la multiplicité est plus efficace que l'estimateur BU (section 4.2);
- ii) considérer l'ajustement par la méthode itérative du quotient (MIO) à des tailles connues de base de sondage N_q , tel qu'il a déjà été proposé pour améliorer l'efficacité de l'estimateur BU (section 4.3);

population inconnues, si bien qu'ils doivent être estimés d'après les données d'échantillon, ce qui complique les calculs et affecte l'optimalité, puisque la variabilité supplémentaire de l'estimation des covariances produit des erreurs quadratiques moyennes plus grandes (Lohr et Rao 2006, section 7).

Afin d'améliorer l'applicabilité, on a proposé d'adopter une approche de *base de sondage unique* (BU) grâce à l'utilisation de *pois fixes* de sorte qu'il n'y ait pas de biais par rapport au plan. Dans le cas de l'échantillonnage aléatoire simple dans chaque base de sondage, l'estimateur BU sur base unique s'obtient en remplaçant dans (2) les poids $w_{(b)}^{f_g}$ par les poids $w_{(b)}^{f_g, \text{SP}} = w_{(K)}^{-1} (\sum_{g \in K} f_g)^{-1}$, où $f_g = n_g / N^g$ désigne la fraction d'échantillonnage dans la base de sondage (Bankier 1986; Kalton et Anderson 1986; Skinner 1991; Skinner, Holmes et Holt 1994). Puisque les poids fixes diffèrent habituellement des poids optimaux, l'estimateur BU est généralement moins efficace que l'estimateur optimal (Lohr et Rao 2000). Enfin, une *approche pseudo-optimal* a été proposée (Skinner et Rao 1996; Lohr et Rao 2000) afin d'obtenir un estimateur d'une applicabilité plus générale que les estimateurs optimaux et d'une plus grande efficacité que les estimateurs BU. Un estimateur du pseudo maximum de vraisemblance (PMV) pour les sondages à bases multiples s'obtient en faisant dans (2) la substitution $w_{(b)}^{f_g, \text{PMV}} = w_{(K)}^{-1} N^K / \sum_{g \in K} \delta_g(K) = N^K / m_K$, où les tailles estimées de domaine N^K sont les solutions d'un système d'équations non linéaires. Bien qu'il soit difficile à mettre en œuvre pour des applications pratiques (une approximation linéaire itérative de N^K sous échantillonnage aléatoire simple est donnée dans Lohr et Rao 2006, section 4.1), l'estimateur PMV garde les bonnes propriétés théoriques de l'approche optimale.

Notons que la formule (2) contient l'indicateur d'appartenance au domaine $\delta_g(K)$; donc les estimateurs optimal, pseudo-optimal et BU s'appliquent uniquement si la classification des données d'échantillon dans les $2^Q - 1$ domaines est effectuée correctement.

À la section suivante, nous présentons un estimateur pour bases de sondage multiples qui s'approche sur une approche à base de sondage unique fondée sur la *multiplicité* ne nécessitant pas de classification par domaine.

3. L'estimateur à base de sondage unique fondé sur la multiplicité

La notion de multiplicité a été introduite pour la première fois dans le contexte de l'échantillonnage par réseau (Casady et Sirken 1980; Sirken 2004). Elle fait aussi partie des outils de la méthode généralisée de partage des poids (Lavallée 2002, 2007), ainsi que de la théorie de l'estimation par échantillonnage de centres (Mecatti 2004).

Puisque, de toute évidence $\sum_{g \in A_g} Y_i = \sum_{i \in U_g} m_i Y_i$, il *de sondage elles appartenant*.

$$(3) \quad Y = \sum_{g=1}^G \sum_{i \in A_g} Y_i m_i^{-1}.$$

Notons que l'expression (3), qui ne comporte que des sommations sur les bases de sondage, offre un avantage pratique par rapport à l'équation (1). En fait, les domaines fournissent une partition virtuelle (inconnue) de la population, tandis que la sélection de l'échantillon a effectivement lieu dans les \bar{Q} bases de sondage chevauchantes. Cela nous mène à un estimateur BU fondé sur la multiplicité donné par

$$(4) \quad Y^M = \sum_{g=1}^G \sum_{i \in A_g} w_{(g)}^{-1} Y_i m_i^{-1} =$$

où les poids fixes $w_{(g)}^{-1}$ permettent de s'assurer, par exemple, de l'absence de biais par rapport au plan. Dans le cas de l'échantillonnage aléatoire simple de chaque base de sondage, nous avons $w_{(g)}^{-1} = \sum_{i \in A_g} Y_i^{-1} \in s_g^q$.

Contrairement aux estimateurs optimal, PMV et BU décrits à la section 2, l'estimateur (4) ne contient pas l'indicateur d'appartenance à l'échantillon et il est très facile à mettre en œuvre dans les applications pratiques. De surcroît, il convient de souligner que, pour l'échantillonnage aléatoire simple de chaque base de sondage, dans l'estimateur fondé sur la multiplicité (4), les valeurs échantillonnées sont pondérées par $(f_g^q m_i^{-1})^{-1}$, c'est-à-dire un coefficient *particulier* à la base de sondage; inversement, dans l'estimateur BU, les valeurs échantillonnées sont pondérées par $w_{(K)}^{-1} = (\sum_{g \in K} f_g^q)^{-1}$, c'est-à-dire un coefficient *moyen* calculé sur l'ensemble des bases de sondage intervenant dans chaque domaine. Par conséquent, l'estimateur Y^M devrait être plus précis que l'estimateur BU, comme le confirment les résultats des simulations. En outre, étant donné sa structure exacte sous une forme explicite. Dans le cas de l'échantillonnage aléatoire simple de chaque base de sondage, la variance de l'estimateur est donnée par

Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples

Fulvia Mecatti¹

Résumé

Les sondages à bases multiples ont été proposés au départ pour favoriser la réduction des coûts dans un contexte d'optimalité. Alors que les sondages de populations *spéciales, rares et difficiles à échantillonner* prennent de l'importance, il arrive souvent, en pratique, que l'on ne dispose pas d'une liste unique des unités de la population comme base de sondage. Récemment, des plans de sondage à bases multiples ont été proposés dans la littérature afin d'accroître la couverture de la population, d'améliorer les taux de réponse et de saisir les différences et les sous-groupes. Diverses approches de l'estimation d'après des bases de sondage multiples ont été présentées, toutes fondées sur la partition virtuelle de l'ensemble de bases de sondage chevauchantes disponibles en domaines disjoints. Par conséquent, la classification correcte des unités d'échantillonnage dans les domaines est requise pour les applications pratiques. Dans le présent article, nous proposons un estimateur pour bases de sondages multiples fondé sur une approche de *multiplicité*. Les estimateurs fondés sur la multiplicité requièrent moins d'information sur l'appartenance d'une unité à un domaine et ne sont donc pas sensibles aux erreurs de classification. En outre, l'estimateur proposé est analytiquement simple, si bien qu'il est facile à appliquer et que sa variance est donnée exactement. Nous présentons aussi les résultats empiriques d'une grande étude par simulation conçue pour comparer l'estimateur fondé sur la multiplicité aux principaux estimateurs concurrents.

Mots clés : Populations difficiles à échantillonner; sondage à base double; erreur de classification; méthode itérative du quotient; estimation de la variance.

1. Introduction

Dans l'échantillonnage en population finie classique, une des hypothèses fondamentales est que l'on dispose comme base d'échantillonnage d'une liste unique et complète des unités formant la population cible. Dans certains cas, un ensemble de deux ou plusieurs listes est disponible pour les besoins du sondage. Le cas général de $\bar{Q} \geq 2$ listes, singulièrement partielles et éventuellement chevauchantes, est appelé *sondage à bases multiples*. Ce type de sondage a été introduit à l'origine (Hartley 1974) comme moyen de réduire les coûts tout en obtenant la même précision que par un sondage à base unique habituel. Dans la pratique contemporaine de l'échantillonnage, les sondages de *populations spéciales, rares et difficiles à échantillonner* deviennent plus fréquents (Kalon et Anderson 1986; Sudman et Kalton 1986; Sudman et Cowan 1988), mais il arrive souvent qu'il n'existe pas de liste unique des unités et que la taille N de la population soit un paramètre inconnu qu'il faut estimer. Les auteurs d'études publiées récemment examinent les sondages à bases multiples dans le but principal d'accroître la couverture de la population, d'améliorer les taux de réponse et de saisir plus exactement les différences et les sous-groupes (Tachan et Dennis 1993; Carlson et Hall 1994; Haines et Pollock 1998; Eurostat 2000). Dans un article récent, Lohr et Rao (2006) mentionnent qu'« à mesure que les populations du Canada, des États-Unis et d'autres pays se

diversifient, l'utilisation de bases de sondage différentes pourrait permettre de mieux refléter les divers sous-groupes. [...] Nous nous attendons à ce que l'usage de plans de sondage modulaires faisant appel à des bases multiples se généralise dans l'avenir » [traduction]. Une application contemporaine de ce principe pourrait s'observer dans les sondages en ligne, dont on peut améliorer la couverture de la population et réduire le biais dû aux caractéristiques du site Internet servant pour la collecte des données en utilisant simultanément deux ou plusieurs sites indépendants. Puisqu'une même unité peut visiter plus d'un site participant au sondage, les sites se chevauchent pour configurer un cadre à bases multiples. L'estimation dans le contexte des sondages à bases multiples, telle qu'elle a été élaborée au départ par Hartley (1962, 1974), est fondée sur la partition virtuelle de la population (c'est-à-dire l'union inconnue des \bar{Q} bases de sondage chevauchantes) en $2^{\bar{Q}} - 1$ *domaines* disjoints (c'est-à-dire les intersections mutuellement exclusives de bases de sondage). Donc, le total X d'une variable étudiée y , considéré comme étant le paramètre à estimer, est exprimé sous forme d'une somme de *totaux de domaine*. Les données d'échantillon provenant des \bar{Q} bases de sondage sont utilisées pour produire des estimations des totaux de domaine. Enfin, les totaux de domaine estimés sont combinés pour produire l'estimation du total de population X . Un certain nombre d'estimateurs ont été élaborés selon diverses approches de l'estimation sur bases

- variables auxiliaires qualitatives parce que les tailles d'échantillon sont faibles ou pour d'autres raisons, les mêmes problèmes de biais que ceux illustrés ici peuvent être introduits dans des estimateurs par calage plus généraux. Un moyen de donner une certaine possibilité de s'écarter des valeurs de contrôle tout en retenant d'importantes variables auxiliaires est déjà proposé dans Rao et Singh (1997). L'effet de leur proposition sur le biais de couverture doit être étudié.
- Remerciements**
- Les auteurs remercient le rédacteur associé et les examinateurs de leur examen minutieux et de leurs commentaires constructifs. Le présent article présente les résultats généraux d'une étude entreprise en partie par le personnel du National Center for Health Statistics (NCHS). Les opinions exprimées n'engagent que les auteurs et ne reflètent pas forcément celles du NCHS. Les travaux de R. Valliant sont financés partiellement en vertu des contrats de services professionnels 200-2004-M-09302 et 200-2006-M-17916 conclus entre l'Université du Michigan et le National Center for Health Statistics.
- Bibliographie**
- Bureau of the Census (2002). *Sources and Accuracy of Estimates for Poverty in the United States: 2002*. Washington DC.
- Eltinge, J., et Yansaneh, I. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.
- Fuller, W. (1966). Estimation employing post strata. *Journal of the American Statistical Association*, 61, 1172-1183.
- Gonzalez, J.F., Town, M. et Kim, J. (2005). Mean square error analysis of health estimates from the Behavioral Risk Factor Surveillance System for counties along the United States/Mexico border region. *Proceedings of the Section on Survey Methods Research*. Alexandria VA: American Statistical Association.
- Kalton, G., et Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.
- Tremblay, V. (1986). Critères pratiques pour la définition des classes de pondération. *Techniques d'enquête*, 12, 91-103.
- Kalton, G., et Maltagliu, D.S. (1991). A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, 409-428.
- Kim, J.J. (2004). Effect of collapsing rows/columns of weighting matrix on weights. *Proceedings of the Section on Survey Methods Research*. American Statistical Association.
- Kim, J.J., Thompson, J.H., Wolman, H.F. et Vais, S.M. (1982). Empirical results from the 1980 Census Sample Estimation Study. *Proceedings of Section on Survey Research Methods*. American Statistical Association, 170-175.
- Kim, J.J., et Tompkins, L. (2007). Comparisons of current and alternative collapsing approaches for improved health estimates. Article présenté à la 11^{ème} Biennial CDC/ASTDR Symposium on Statistical Methods, Atlanta, Georgia, 17-18 avril 2007.
- Kim, J.J., Tompkins, L., Li, J. et Valliant, R. (2005). A simulation study of cell collapsing in poststratification. *Proceedings of the Section on Survey Methods Research*. American Statistical Association.
- Kostanich, D., et Dippo, C. (2000). *Current Population Survey: Design and Methodology*. Article technique 63. Washington DC: Department of Commerce.
- Lazzeroni, L., et Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, 14, 61-78.
- Little, R.J.A. (1993). Post-stratification: A modeler's perspective. *Journal of American Statistical Association*, 88, 1001-1012.
- Little, R.J.A., et Vartivartan, S. (2005). La pondération pour la non-réponse augmente-t-elle la variance des moyennes de sondage? *Techniques d'enquête*, 31, 175-183.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9, 1-19.
- Lumley, T. (2005). Survey: Analysis of complex survey samples. R package version 3.0-1. University of Washington: Seattle.
- R Development Core Team (2005). *R: A language and environment for Statistical Computing*. <http://www.R-project.org>.
- Rao, J.N.K., et Singh, A.C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. *Proceedings of the Survey Research Methods Section*. Alexandria VA: American Statistical Association.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Tompkins, L., et Kim, J.J. (2006). Evaluation of collapsing criteria in sample weighting. Internal NCHS memorandum.

Tableau 6 Ratios des variances (ou des EQM) à la variance (ou à l'EQM) de l'estimateur poststratifié sans regroupement de cellules (PS1) Les chiffres pour Hájek sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Ratios des variances à la variance de l'estimateur poststratifié (PS1)		Ratios des EQM à l'EQM de l'estimateur poststratifié (PS1)	
Hájek	PS2	PS.WR1	Hájek	PS2
(regroupement (troncature) des poids puits maximale fixe standard)		(regroupement (troncature) des poids puits maximale fixe standard)		PS.WR2
(regroupement) des poids)		(regroupement) des poids puits maximale fixe standard)		PS.WR2
(regroupement) des poids)		(regroupement) des poids puits maximale fixe standard)		PS.WR2

Assurance-maladie	0,877	1,014	1,025	0,991	1,500	1,101	1,018	1,006
Limitations	0,821	1,035	0,977	1,555	1,008	1,017	1,006	1,006
Retard des soins	1,099	1,003	1,000	1,239	1,001	1,000	1,000	1,000
Hospitalisation	1,290	1,011	1,000	1,733	1,008	1,012	1,000	1,012
Moyenne commune Y	0,776	0,935	0,974	0,902	0,781	0,933	0,973	0,906
Regroupement par adjacence, seuil de correction = 1,8								
Assurance-maladie	0,877	0,960	1,044	0,976	1,500	1,179	1,024	1,048
Limitations	0,821	0,939	1,032	0,961	1,555	1,034	1,017	1,000
Retard des soins	1,099	1,051	1,025	1,091	1,239	1,057	1,034	0,989
Hospitalisation	1,290	1,225	1,043	1,201	1,733	1,442	1,023	1,256
Moyenne commune Y	0,780	0,815	0,882	0,828	0,779	0,816	0,893	0,829
Regroupement selon la moyenne proche, seuil de correction = 1,8								
Assurance-maladie	0,877	1,010	1,006	1,019	1,500	1,018	1,000	1,024
Limitations	0,821	0,983	1,051	0,975	1,555	1,034	1,017	1,000
Retard des soins	1,099	1,003	0,995	1,001	1,239	1,035	1,000	1,000
Hospitalisation	1,290	1,052	1,001	1,059	1,733	1,032	1,000	1,047
Moyenne commune Y	0,771	0,924	0,958	0,876	0,778	0,932	0,959	0,879

Tableau 7 Taux de couverture en pourcentage des intervalles de confiance à 95 % calculés en utilisant la loi t avec 25 DDL (Les chiffres pour Hájek et PS1 ne sont pas affectés par le regroupement et sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Hájek	PS1	PS2	PS.WR1	PS.WR2
	(pas de regroupement)	(regroupement standard)	(troncature des poids puits regroupement)	(correction maximale fixe des poids)	

La partie de droite du tableau 6 donne les ratios des erreurs quadratiques moyennes (EQM) empiriques des proportions estimées, exprimées en pourcentage de l'EQM de PS1. À quelques exceptions près, PS2 est l'estimateur poststratifié donnant les pires résultats pour les quatre premières caractéristiques, quelle que soit la combinaison de variable, $f_{\max} = 1,8$, c'est-à-dire le choix dominant lieu à un plus grand nombre de regroupements, les EQM de PS2 excède de 1,8 % à 44,2 % celles de PS1. Les EQM de PS.WR1 ainsi que PS.WR2 sont proches de celles de PS1, à l'exception du cas (hospitalisation, $f_{\max} = 1,8$, adjacence) où le biais de 6,3 % de PS.WR2 donne lieu à une EQM

De nouveau, les estimateurs donnent des résultats différents en ce qui concerne la variable de moyenne commune Y. Les EQM de Hâjek, PS2, PS.WR1 et PS.WR2 sont toutes inférieures à celles de PS1. L'estimateur de Hâjek est celui dont l'EQM est la plus faible, parce que la poststratification n'est pas nécessaire pour corriger le biais dans l'estimation de la moyenne.

Tableau 4 Erreur absolue moyenne de correction de la couverture telle qu'elle est définie dans l'expression (10)

Méthode de regroupement	Seuil de correction	Hâjek	PS2	PS.WR1	PS.WR2
f_{\max}			(regroupement standard)	(troncature des poids puis regroupement)	(troncature des poids puis regroupement)
C1 Ratios de couverture pour la variable d'assurance-maladie					
Adjacence	2	0,257	0,120	0,086	0,221
Moyenne proche	2	0,257	0,080	0,127	0,281
Adjacence	1,8	0,256	0,150	0,085	0,020
Moyenne proche	1,8	0,256	0,101	0,109	0,258
C5 Ratio de couverture pour la variable de moyenne commune Y					
Adjacence	2	0,442	0,326	0,196	0,331
Moyenne proche	2	0,441	0,321	0,203	0,370
Adjacence	1,8	0,442	0,330	0,206	0,376
Moyenne proche	1,8	0,442	0,337	0,214	0,446

Tableau 5 Biases relatifs (en pourcentage) des proportions estimées (Les chiffres pour Hâjek et PS1 ne sont pas affectés par le regroupement et sont répétés dans les quatre parties du tableau pour faciliter les comparaisons)

Caractéristiques	Fourchette du nombre de post-strates après le regroupement	Hâjek	PS1 (pas de regroupement)	PS2 (regroupement standard)	PS.WR1 (troncature des poids puis regroupement)	PS.WR2 (troncature des poids puis regroupement)

Assurance-maladie	(10, 16)	-11,5	0,1	-4,4	1,0	-1,4
Limitations	(8, 15)	-12,1	-0,3	-2,0	0,1	-1,0
Retard des soins	(6, 14)	8,2	-0,2	2,2	-0,6	0,9
Hospitalisation	(9, 16)	13,4	0,0	6,2	-0,7	2,8
Moyenne commune Y	(5, 11)	0,3	0,0	0,4	0,4	0,6
Regroupement selon la moyenne proche, seuil de correction = 2						
Assurance-maladie	(10, 16)	-11,5	0,1	-0,5	0,5	-0,3
Limitations	(8, 15)	-12,1	-0,3	-1,2	0,2	-1,1
Retard des soins	(6, 14)	8,2	-0,2	-0,3	-0,3	-0,2
Hospitalisation	(9, 16)	13,4	0,2	0,4	0,1	0,4
Moyenne commune Y	(6, 11)	0,3	0,0	0,2	0,1	0,2
Regroupement par adjacence, seuil de correction = 1,8						
Assurance-maladie	(7, 13)	-11,5	0,1	-6,5	0,7	-3,5
Limitations	(7, 12)	-12,1	-0,3	-3,4	0,3	-2,0
Retard des soins	(5, 11)	8,2	-0,2	3,5	-0,4	2,5
Hospitalisation	(5, 12)	13,4	0,2	9,4	0,0	6,3
Moyenne commune Y	(5, 9)	0,3	0,1	0,5	0,6	0,6
Regroupement selon la moyenne proche, seuil de correction = 1,8						
Assurance-maladie	(6, 13)	-11,5	0,1	-1,6	0,3	-1,7
Limitations	(7, 12)	-12,1	-0,3	-2,7	0,9	-2,4
Retard des soins	(5, 10)	8,2	-0,2	0,2	-0,3	0,5
Hospitalisation	(5, 12)	13,4	0,2	1,5	0,3	2,0
Moyenne commune Y	(5, 10)	0,3	0,2	0,3	0,3	0,3

Dans les cas de PS2, PS.WR1 et PS.WR2, ces estimateurs de variance ne tiennent pas compte de la nature dynamique du regroupement de cellules qui peut varier d'un échantillon à l'autre. Par conséquent, l'une des sources de variation n'est pas prise en compte et nous pouvons nous attendre à ce que les estimations de la variance soient un peu trop faibles comparativement aux variances empiriques obtenues par simulation.

4.6 Résultats des simulations

Les tableaux 4 à 7 résument les résultats pour les erreurs de correction de la couverture, les biais relatifs des proportions estimées, les variances des divers estimateurs étudiés et la couverture des intervalles de confiance calculée en utilisant les estimateurs de variance par linéarisation. Le tableau 4 donne l'erreur absolue moyenne de correction de la couverture, définie comme étant

$$\bar{e} = (D)^{-1} \sum_{i=1}^D \left| N_i^w / N_i - 1 \right| \quad (10)$$

où d est l'un des $D=2000$ échantillons et N_i^w est le nombre estimé d'unités dans la post-strate i en se basant sur les poids finaux pour un estimateur particulier (Hajék, PS2, PS.WR1 ou PS.WR2). La valeur de \bar{e} est 0 pour l'estimateur poststratifié sans regroupement de cellules, PS1, puisqu'il corrige entièrement l'erreur de couverture dans chacune des 16 post-strates. Pour illustrer comment les erreurs moyennes de correction de la couverture peuvent varier, nous avons estimé les proportions pour les variables d'assurance-maladie et de moyenne commune Y en utilisant les ratios de couverture de la base de sondage C1 et C5. Pour la plupart des combinaisons de ratio de couverture, méthode de regroupement et seuil de correction, PS.WR1 corrige plus efficacement l'erreur de couverture que l'estimateur par regroupement standard, PS2. Par exemple, $\bar{e} = 0,086$ dans le cas de PS.WR1 pour (assurance-maladie, regroupement par adjacence, $f_{\max} = 2$), tandis que $\bar{e} = 0,120$ dans le cas de PS2. En revanche, la correction de la couverture est un peu moins bonne si l'on utilise PS.WR2

Le tableau 5 donne les biais relatifs, définis comme étant $100 \sum_{d=1}^D \left(\frac{\bar{y}_d}{Y} - \frac{\bar{y}_d}{Y} \right)$, où \bar{y}_d est l'une des estimations de la proportion pour l'échantillon d . Les estimations de Hajék sont fortement biaisées pour les quatre premières caractéristiques, puisqu'elles ne comportent aucune correction pour les différences de sous-couverture entre les cellules. Les biais relatifs varient de -12,1 % pour les limitations fonctionnelles à 13,4 % pour l'hospitalisation. Comme nous l'avons mentionné à la section 2, le biais peut être positif ou négatif, en fonction de la corrélation des taux de couverture et des moyennes de cellule.

La poststratification sans regroupement de cellules (PS1) produit des estimations presque sans biais, tandis que les autres options - PS2, PS.WR1 et PS.WR2 - introduisent toutes un biais lorsqu'on utilise le regroupement par adjacence pour les quatre premières caractéristiques. Le nombre de post-strates après le regroupement, donné au tableau 5, varie de 6 à 16 quand $f_{\max} = 2$ et de 5 à 13 quand $f_{\max} = 1,8$. Le biais relatif de PS2, si l'on utilise le regroupement par adjacence, varie de -4,4 % à 6,2 % quand $f_{\max} = 2$ et de -6,5 % à 9,4 % quand $f_{\max} = 1,8$. Dans le cas du regroupement par adjacence, les options de rechange, PS.WR1 et PS.WR2, produisent des biais compris entre ceux de PS1 (aucun regroupement) et de PS2. En particulier, PS.WR1 donne des résultats raisonnablement comparables à ceux de PS1 en ce qui concerne le biais avec regroupement par adjacence. En revanche, le regroupement selon la moyenne proche produit des estimations PS2, $f_{\max} = 2$. Dans le cas du regroupement selon la moyenne et $f_{\max} = 1,8$, les estimateurs PS2 et PS.WR2 présentent encore un léger biais, mais PS.WR1 se compare bien à PS1. Pour la cinquième caractéristique (moyenne commune Y), tel que prévu, tous les estimateurs sont presque sans biais, quelle que soit la méthode de regroupement des cellules.

L'une des justifications conventionnelles du regroupement des cellules est que les poids extrêmes seront réduits et que les variances des estimations seront, à leur tour, plus faibles. Le tableau 6 donne les ratios des variances empiriques des proportions estimées par rapport à la variance de PS1. L'estimateur de Hajék donne des estimations dont les variances sont de 12 % à 18 % plus faibles que celles des estimations produites par PS1 pour l'assurance-maladie et les limitations fonctionnelles, mais qui sont plus variables que ces dernières pour le retard des soins médicaux et l'hospitalisation. Ces résultats montrent, en outre, clairement la variance d'un estimateur poststratifié. L'utilisation de PS2 produit de légères améliorations de la variance pour certaines combinaisons des quatre premières variables, mais sous le scénario (adjacence, $f_{\max} = 2$), la variance de PS2 pour l'hospitalisation est 17 % plus grande que si l'on utilise PS1. Dans le cas de (adjacence, $f_{\max} = 1,8$), PS2 donne des estimations 23 % plus variables pour l'hospitalisation. PS.WR1 ne produit pas de variances extrêmes de PS2 dans le cas du regroupement par adjacence; comme PS2, PS.WR2 possède une plus grande variance pour l'hospitalisation sous regroupement par adjacence. Si l'on recourt au regroupement selon la moyenne proche plutôt que l'adjacence, les variances de PS2, PS.WR1 et PS.WR2 sont sensiblement plus proches de celles de PS1. Toutefois, pour la variable de moyenne commune Y , le regroupement réduit systématiquement la variance. La réduction est de près de 20 % pour le regroupement par adjacence.

	Age	< 5 ans	5 à 17 ans	18 à 24 ans	25 à 44 ans	45 à 64 ans	65 à 69 ans	70 à 74 ans	75 ans et plus
C1: Non couvert-maladie	Femmes	0,9	0,8	0,5	0,8	0,9	0,9	0,9	0,9
C1: Non couvert-par un assurance-maladie	Femmes	0,9	0,8	0,5	0,8	0,9	0,9	0,9	0,9
C2: Limitations physiques, mentales, émotionnelles	Femmes	0,9	0,6	0,8	0,5	0,5	0,5	0,5	0,5
C3: Retard des soins médicaux au cours des 12 derniers mois	Femmes	0,5	0,5	0,8	0,8	0,5	0,5	0,5	0,5
C4: Hospitalisation au cours des 12 derniers mois	Femmes	0,8	0,5	0,8	0,5	0,8	0,8	0,8	0,8
C5: Moyenne commune Y	Femmes	0,8	0,7	0,4	0,6	0,3	0,2	0,4	0,9

Pour chacun des estimateurs de proportion, nous avons calculé une estimation de la variance par linéarisation. Chaque estimation de la variance est basée sur la méthode de substitution linéaire (par exemple, voir Samdal et coll. 1992, chapitre 5). Nous avons calculé des estimations de la variance pour tous les estimateurs de proportion en utilisant les fonctions `svydesign`, `poststratify` et `svymeans` du logiciel pour données d'enquête R. L'approche théorique générale consiste à faire une approximation linéaire d'un estimateur particulier. L'approximation linéaire est réarrangée de façon que l'estimateur soit écrit sous la forme d'une somme de totaux pondérés d'UPE et l'estimateur de variance utilisé est celui applicable à l'échantillonnage d'UPE avec remise. Les estimateurs \hat{Y}_{PS1} , \hat{Y}_{PS2} et \hat{Y}_{PSWR2} sont traités comme des estimateurs poststratifiés standard pour les besoins de

Chiffres de population		Non couverte par une assurance-maladie		Limitations physiques, mentales, émotionnelles		Retard des soins médicaux au cours des 12 derniers mois		Hospitalisation au cours des 12 derniers mois		Total	
Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes	Hommes	Femmes
843	795	1 638	1 010	3	4	3	3	3	4	16	15
2 271	2 082	4 353	13	10	6	4	8	4	4	2	2
2 971	3 207	6 178	28	25	7	9	8	10	3	6	8
2 211	2 597	5 018	14	16	19	7	11	9	9	10	9
305	384	689	2	24	29	3	8	6	15	14	14
423	717	1 140	1	41	48	2	5	4	22	22	22
10 507	11 157	21 664	18	16	13	6	8	7	7	10	8
Total											

Tableau 2 Pourcentage de personnes possédant quatre caractéristiques liées à la santé dans les groupes formés selon l'âge et le sexe

Dans le cas de la méthode d'adjacence, une cellule à faible effectif était regroupée à la voisine ayant le plus faible facteur de poststratification. Dans le cas de la méthode de la moyenne proche, une cellule à faible effectif était regroupée avec l'échantillon non pondérée était la plus proche de la sienne.

1	5
2	6
3	7
4	8

Dans le regroupement fondé sur l'adjacence, les voisines « adjacentes » et « moyenne proche ».

Nous avons établi des situations où les conditions pour l'absence de biais décrites aux sections 2 et 3 pouvaient être violées lorsque des cellules étaient regroupées dans les simulations. Chacun des estimateurs, $\hat{Y}_{PS,WR1}^{PS2}$ et $\hat{Y}_{PS,WR2}^{PS2}$, comporte un regroupement de cellules. Si le FCI (facteur de poststratification) dans une post-strate initiale, N_i/N_{zi} est suspecteur à la correction maximale permise, f_{max} , ou que la taille de l'échantillon de cellule est inférieure à un minimum fixé, n_{min} , nous disons que cette post-strate est une cellule « à faible effectif » et nous la regroupons avec une cellule voisine. Nous avons utilisé deux méthodes de sélection des voisines, appelées ici « adjacence » et « moyenne proche ».

4.4 Règles de regroupement

population de cette post-strate ont été sélectionnés aléatoirement pour demeurer dans la base de sondage, tandis que pour le reste de cette population, la probabilité d'être échantillonné était nulle.

Cinq ensembles de scénarios de couverture, présentés au tableau 3, ont été utilisés pour filtrer la population avant d'échantillonner les UPE. Le ratio de couverture variait selon la post-strate et différait pour chacune des cinq caractéristiques pour lesquelles les proportions ont été estimées. Les ratios de couverture particuliers à chacune des cinq caractéristiques sont nommés de C1 à C5 dans le tableau 3. Ils ont été créés artificiellement d'après les moyennes de population pour chaque groupe âge-sexe. Une moins bonne couverture a été attribuée au groupe comportant une portion plus élevée d'unités possédant les caractéristiques de couverture par une assurance-maladie et de limitations fonctionnelles; la situation était inverse pour le retard des soins médicaux et l'hospitalisation. Dans le scénario C5, les ratios de couverture sont assez différents afin de produire des corrections de la couverture qui varient sensiblement parmi dans l'ensemble initial de 16 post-strates. Quoique les taux présentés au tableau 3 soient faibles, ils sont comparables à ceux donnés pour la BRFS au tableau 1, ou plus élevés que ceux-ci. En appliquant ces taux, nous avons sélectionné aléatoirement un sous-ensemble de la population devant figurer dans la base de sondage pour chaque échantillon qui avait été sélectionné. Par exemple, si le ratio de couverture dans la post-strate des garçons de moins de 5 ans était de 0,9, 90 % de la

4.3 Scénarios de couverture

les estimateurs des proportions de population décrits aux sections 2 et 3, c'est-à-dire l'estimateur de Hajék, \hat{Y}_{pi} , l'estimateur poststratifié \hat{Y}_{PSi}^{PS1} , qui utilise l'ensemble des 16 post-strates, l'estimateur poststratifié avec regroupement de cellules, $\hat{Y}_{PS,WR2}^{PS2}$, et les deux estimateurs à poids restreint $\hat{Y}_{PS,WR1}^{PS2}$ et $\hat{Y}_{PS,WR2}^{PS2}$. Le code de simulation a été rédigé dans le langage R (R Development Core Team 2005) en servant intensivement du progiciel pour données d'enquête (Lumley 2004, 2005).

limitations physiques, mentales ou émotionnelles, c'est-à-dire si une personne était limitée ou non de n'importe laquelle de ces façons; retard des soins médicaux, c'est-à-dire si une personne avait retardé ou non des soins médicaux à cause de leur coût au cours des 12 derniers mois; séjour d'au moins une nuit à l'hôpital, c'est-à-dire si une personne avait passé ou non une nuit à l'hôpital au cours des 12 derniers mois.

où $q_i = 1 - p_i$, $(\overline{q})^g = \sum q_i p_i / N^g$, et où $C^{p_i, g}$ ont été définis antérieurement. Ensuite, utilisons le fait que $\sum q_i N_i^g = N^g - N^g$ pour définir $p^g = N^g / N^g$ la proportion d'unités couvertes dans le groupe g , et $p^g = N^g / N^g$ la proportion d'unités couvertes prévue dans les cellules à faible effectif dans le groupe g . Alors, le biais peut aussi s'écrire sous la forme

$$\text{biais}(\hat{p}_{PS, WR2}^g) = \frac{1}{I} \sum_{g=1}^G f_{\max}^g (N^g (\mu_{g, sp}^g - \mu_{g, sp}^g)) + \sum_{g=1}^G W^g \left(C^{\psi_g, g} - C^{\psi_g, g} - \frac{p^g}{p^g} \right) \quad (8)$$

À en juger par l'expression (8), $\hat{p}_{PS, WR2}^g$ sera approximativement sans biais si la moyenne par unité pour les unités couvertes par la base de sondage dans chaque cellule regroupée est la même dans les cellules à faible effectif, $\mu_{g, sp}^g$, que dans les cellules à effectif non faible, c'est-à-dire $\mu_{g, sp}^g = \mu_{g, sp}^g$, et que les covariances, $C^{\psi_g, g}$ et $C^{\psi_g, g}$, sont toutes deux nulles. La dernière condition est réalisée en combinant des cellules ayant les mêmes moyennes, \bar{Y}_i . Le fait de combiner des cellules dont les taux de couverture sont égaux ne produira pas un $\hat{p}_{PS, WR2}^g$ sans biais. La situation est plus limitée que pour $\hat{p}_{PS, WR1}^g$, qui est sans biais si les taux de couverture ou les moyennes sont les mêmes dans toutes les cellules d'un groupement.

4. Une études empirique

Après d'évaluer certaines idées présentées plus haut, nous avons réalisé une étude par simulation des propriétés de biais des diverses méthodes de poststratification proposées. Nous avons aussi examiné les propriétés d'un estimateur de variance qui est souvent utilisé en pratique.

4.1 Population étudiée

La population utilisée dans la simulation a été extraite du fichier de microdonnées à grande diffusion de la National Health Interview Survey (NHIS) de 2003. Nous avons créé un sous-ensemble du fichier de la NHIS contenant 21 664 personnes. Les strates et les UPE étaient fondées sur celles du fichier de données à grande diffusion de la NHIS, mais nous avons regroupé les strates par trois pour créer de nouvelles strates d'échantillonnage pour la population étudiée. Nous avons utilisé quatre variables binaires (caractéristiques ayant une valeur 0 ou 1) pour la simulation, fondées chacune sur une réponse autodéclarée par la personne, soit :

couverte par une assurance-maladie, c'est-à-dire si une personne était couverte ou non par tout type d'assurance-maladie;

Le tableau 2 donne les pourcentages de personnes présentant ces quatre caractéristiques dans les cellules formées par l'âge et le sexe. Ces 16 cellules (âge \times sexe) représentent l'ensemble initial de post-strates utilisées pour l'estimation. Les pourcentages peuvent varier considérablement d'une cellule à l'autre, selon les caractéristiques. Ainsi, les personnes de 18 à 24 ans sont nettement plus susceptibles que les autres de ne pas être couvertes par une assurance-maladie, tandis que les enfants de moins de cinq ans et les personnes de 65 ans et plus sont beaucoup plus susceptibles que les autres d'avoir été hospitalisés. Le regroupement de cellules dont les moyennes, ou les proportions dans ce cas-ci, diffèrent peut introduire un biais, comme nous l'avons mentionné plus haut. Nous avons également créé une variable binaire artificielle ayant une moyenne commune de 0,20, quelle que soit l'appartenance de l'unité à la post-strate. Dans ces conditions, tous les estimateurs, y compris celui de Hajek, seront sans biais indépendamment des taux de couverture. En outre, l'idée conventionnelle selon laquelle le regroupement de cellules peut réduire les variances en lissant les corrections extrêmes des poids pourrait s'avérer vraie pour cette variable.

4.2 Plan d'échantillonnage

Deux UPE ont été sélectionnées dans chaque strate avec probabilité proportionnelle à la taille (PPT), la taille étant le nombre de personnes dans chaque UPE. L'échantillonnage des UPE a été fait avec remise afin de simplifier l'estimation de la variance. Si nous avions utilisé un échantillonnage sans remise, une méthode plus élaborée de sélection et d'estimation de la variance aurait été nécessaire (voir, par exemple, Samdal, Swensson et Wretman 1992, chapitre 3). Dans chaque UPE échantillonnée, 20 personnes ont été sélectionnées par échantillonnage aléatoire simple sans remise pour donner un total de 1 000 personnes dans chaque échantillon. Pour chaque combinaison de paramètres discutée plus loin, nous avons sélectionné 2 000 échantillons.

Nous avons utilisé 16 post-strates initiales obtenues par le croisement des huit groupes d'âge présentés au tableau 2 avec le sexe. Dans chaque échantillon, nous avons calculé

L'algorithme détaillé de calcul des poids pour l'option PS.WR2 est le suivant :

(1) Exécuter les étapes (1) à (3) de l'algorithme de la section 2.3 pour PS2.

(2) Dans un groupe contenant au moins une cellule à effectif non faible, calculer le total de contrôle dans le groupe g comme étant $N_g^* = \sum_{i \in A_g^*} N_i$ et le poids corrigé pour toutes les unités k dans $A_g^{*,sp}$ comme étant $w_k = w_k f_{\max}^*$.

(3) Calculer le poids corrigé pour toutes les unités k dans $A_g^{*,sp}$ comme étant $w_k(N_g^* - \hat{N}_{g,sp}^*) / (\hat{N}_{g,sp}^* / N_{g,sp}^*)$ où $N_{g,sp}^* = \sum_{i \in A_g^{*,sp}} w_k$ et $\hat{N}_{g,sp}^* = \sum_{h,j \in s_h, k \in s_{hj(i)}} w_k$.

(4) Le poids corrigé final est alors w_k pour l'unité k dans le groupe g .

Ce deuxième estimateur à poids restreints peut s'écrire sous la forme $\hat{Y}_{PS,WR2}^g = (1/N) \sum_{i \in A_g^*} T_i$ où

$$T_i = \sum_{h,j \in s_h, k \in s_{hj(i)}} f_{\max} w_k Y_k$$

$$+ \sum_{i \in A_g^{*,sp}} \sum_{h,j \in s_h, k \in s_{hj(i)}} \frac{N_{g,sp}^*}{N_g^* - f_{\max} \hat{N}_{g,sp}^*} w_k Y_k$$

où $\hat{N}_{g,sp}^* = \sum_{i \in A_g^{*,sp}} \sum_{h,j \in s_h, k \in s_{hj(i)}} w_k$. L'espérance de T_i en fonction de la couverture, du degré de faible effectif et du plan d'échantillonnage est

$$E(T_i) = f_{\max} \frac{N_g^*}{N_g^* - f_{\max} \hat{N}_{g,sp}^*} (T_g^c - \bar{T}_g^c) + \frac{N_g^*}{N_g^* - f_{\max} \hat{N}_{g,sp}^*} (T_g^c - \bar{T}_g^c)$$

où $\bar{T}_g^c = \sum_{i \in A_g^*} T_i / N_g^*$. Après certains calculs, le biais approximatif de $\hat{Y}_{PS,WR2}^g$ peut s'écrire

$$\text{biais}(\hat{Y}_{PS,WR2}^g) = \frac{1}{N} f_{\max} \sum_{i \in A_g^*} \bar{N}_i^c (\mu_{i,sp}^c - \mu_{i,sp}^*) + \frac{\sum_{i \in A_g^*} w_i^* \bar{N}_i^c}{1} \times$$

$$\left[\sum_{i \in A_g^*} q_i T_i^c - N_{g,sp}^* \left(\sum_{i \in A_g^*} q_i N_i^* \right) \left(\sum_{i \in A_g^*} T_i \right) \right]$$

où $\mu_{i,sp}^c = \bar{T}_i^c / \bar{N}_i^c$ et $\mu_{i,sp}^* = \sum_{i \in A_g^*} q_i T_i^c / \sum_{i \in A_g^*} q_i N_i^*$. Maintenant, notons que dans le cas d'une probabilité de couverture commune dans la cellule i , $\phi_i = \phi(i)$,

$$\sum_{i \in A_g^*} q_i T_i^c - N_{g,sp}^* \left(\sum_{i \in A_g^*} q_i N_i^* \right) \left(\sum_{i \in A_g^*} T_i \right) = \sum_{i \in A_g^*} \sum_{h,j \in U_h, k \in U_{hj(i)}} (q_i \phi_k - (q_i \phi)(i)) (T_i^c - \bar{T}_g^c) = \sum_{i \in A_g^*} N_i^* q_i \phi(i) (T_i^c - \bar{T}_g^c) = \sum_{i \in A_g^*} N_i^* (C_i^* - C^{p\phi, Y, g})$$

diffèrent. Il est presque certain que cette condition ne sera pas vérifiée aussi longtemps qu'un post-strate i d'un groupe présente une probabilité d'avoir un faible effectif sensiblement différente de celle des autres.

Dans le cas d'une probabilité de couverture commune dans la post-strate i , $\phi(i)$, nous pouvons aussi comparer le biais de l'estimateur à cellules regroupées, \hat{Y}_{PS2}^g , à celui de $\hat{Y}_{PS,WR1}^g$. En utilisant les résultats du paragraphe précédent, dans (6), le biais peut être exprimé comme étant

$$\text{biais}(\hat{Y}_{PS,WR1}^g) = \sum_{i \in A_g^*} W_i^* \left[\frac{C_i^{*\phi} / \phi_i + (f_{\max}^* - 1) C^{p\phi, Y, g} / \phi_g}{1 + (f_{\max}^* - 1) (p)_{\phi_g}^g / \phi_g} - 1 \right] (p)_{\phi_g}^g / \phi_g - 1 + (f_{\max}^* - 1) (p)_{\phi_g}^g / \phi_g \geq 1, \text{ nous pouvons utiliser}$$

(5) pour obtenir

$$\left| \text{biais}(\hat{Y}_{PS,WR1}^g) \right| \leq \left| \sum_{i \in A_g^*} W_i^* C_i^{*\phi} / \phi_i + (f_{\max}^* - 1) C^{p\phi, Y, g} / \phi_g \right| = \left| \text{biais}(\hat{Y}_{PS2}^g) + (f_{\max}^* - 1) \sum_{i \in A_g^*} W_i^* C^{p\phi, Y, g} / \phi_g \right|.$$

Si $p_i \phi(i)$ et \bar{Y}_g ne sont pas corrélés, le biais absolu de $\hat{Y}_{PS,WR1}^g$ est inférieur ou égal à celui de \hat{Y}_{PS2}^g , parce que $1 + (f_{\max}^* - 1) (p)_{\phi_g}^g / \phi_g \geq 1$. Si $p_i \phi(i)$ et \bar{Y}_g sont corrélés, il faut considérer deux cas : i) biais $(\hat{Y}_{PS2}^g) \geq 0$ et ii) biais $(\hat{Y}_{PS2}^g) < 0$. Dans le premier, la dernière ligne de (7) sera inférieure ou égale au biais absolu de \hat{Y}_{PS2}^g si

$$\left| -2 \left| \text{biais}(\hat{Y}_{PS2}^g) \right| \frac{f_{\max}^* - 1}{f_{\max}^*} \right| \leq \sum_{i \in A_g^*} W_i^* C^{p\phi, Y, g} / \phi_g \leq 0.$$

Dans le cas ii), l'exigence est que

$$0 \leq \sum_{i \in A_g^*} W_i^* C^{p\phi, Y, g} / \phi_g \leq 2 \left| \text{biais}(\hat{Y}_{PS2}^g) \right| \frac{f_{\max}^*}{f_{\max}^* - 1}.$$

Si la covariance entre les probabilités d'avoir un faible effectif et d'être couvert, $p_i \phi(i)$, et les moyennes de cellules, \bar{Y}_g , est faible dans tous les groupes et de signe opposé au biais (\hat{Y}_{PS2}^g) , alors $\hat{Y}_{PS,WR1}^g$ aura un biais plus faible que \hat{Y}_{PS2}^g .

La deuxième option, dénotée PS.WR2, a pour but d'exercer un plus grand contrôle sur la grandeur de la correction finale des poids que ne le fait l'option PS.WR1. Dans le cas de cette dernière, la correction finale peut être plus grande que f_{\max}^* . L'option PS.WR2 vise à limiter la correction finale à $f_{\max}^* = 2$ ou à un autre maximum fixe d'avance. L'idée générale est de déterminer d'abord quelles cellules devraient être regroupées, comme dans le cas de PS.WR1. Ensuite, les poids dans les cellules à faible effectif sont multipliés par f_{\max}^* . Ceux dans la cellule à effectif non faible d'un regroupement sont ajustés par un facteur constant afin de faire concorder le chiffre estimé de population dans le groupe au dénombrement de contrôle.

Donc, dans le cas où \hat{y}^{ps1} est sans biais, \hat{y}^{ps2} sera biaisé si l'on regroupe des post-strates dont les taux de couverture et les moyennes de population sont différents. Puisque $\hat{\phi}^g$ et C^g sont $O(1)$, le biais ne diminue pas à mesure que l'échantillon s'accroît; donc, le carré du biais finira par être la partie dominante de l'erreur quadratique moyenne. Si les cellules sont regroupées, celles qui constituent un même groupe devraient avoir les mêmes taux de couverture, les mêmes moyennes, ou les deux, afin d'éviter le biais.

3. Estimateurs à poids restreints

Nous examinons deux méthodes alternatives de calcul de la pondération en cas de regroupement de post-strates qui s'inscrivent dans la foulée des travaux de Kim (2004). Les deux options sont conçues de façon à obtenir un compromis entre a) l'utilisation de toutes les post-strates et l'éventualité de post-strates produisant des poids moins variables, mais éventuellement des estimations biaisées. Nous les appelons méthodes de *restriction des poids* (WR pour *weight restriction*). Les deux options décrites à la présente section comportent le regroupement de cellules, mais conservent une part plus importante de la correction des poids des cellules individuelles que la méthode standard de regroupement de cellules.

La première option, que nous dénotons PS.WR1, consiste en l'algorithme suivant. Designons la correction maximale permise des poids par f_{\max} , avec $f_{\max} > 1$.

1) Exécuter les étapes 1) à 3) de l'algorithme de la section 2.3 pour PS2.

2) Censurer tout FCI plus grand que f_{\max} à f_{\max} et ajuster chaque poids dans la cellule initiale correspondante à $w_k = w_k / f_{\max}$ avec $w_k = 1/\pi_k$. Pour les unités comprises dans les cellules pour lesquelles $FCI \leq f_{\max}$, fixer $w_k = w_k$.

3) Calculer un facteur de correction par regroupement (FCR) pour un groupe g sous la forme

$$\bar{f}^g = N^g / \sum_{i \in A_{i,sp}^g} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k.$$

4) Le poids corrigé final est alors $w_k \bar{f}^g$ pour l'unité k dans le groupe g .

Cette méthode réduira les valeurs les plus grandes de la correction finale des poids de façon à les rendre inférieures aux corrections sans regroupement de cellules, N^g/N^{tr} , quoiqu'il puisse exister un ou plusieurs regroupements pour lesquels le FCR est plus grand que le seuil f_{\max} . Le total de contrôle pour le groupe g , N^g , est respecté en ce sens que $\sum_{i \in A_{i,sp}^g} \sum_{h, j \in s_h, k \in s_{hj(i)}} w_k \bar{f}^g = N^g$, mais les totaux de contrôle

Pour analyser les propriétés de PS.WR1, définissons $A_{i,sp}^g$ et $A_{i,sp}^{g,sp}$ comme étant les ensembles de post-strates à effetif faible et à effetif non faible dans le groupe de cellules g . PS.WR1 peut être exprimé par

$$\hat{y}^{ps,WR1} = \frac{1}{N^g} \sum_{i \in A_{i,sp}^g} \hat{y}_i^{ps,WR1} = \frac{1}{N^g} \sum_{i \in A_{i,sp}^g} \sum_{h, j \in s_h} \sum_{k \in s_{hj(i)}} w_k y_k.$$

où

$$\hat{y}_i^{ps,WR1} = \sum_{h, j \in s_h} \sum_{k \in s_{hj(i)}} \sum_{i' \in A_{i',sp}^g} w_{k'} y_{k'} + \sum_{i' \in A_{i',sp}^g} \sum_{h, j \in s_h} \sum_{k \in s_{hj(i)}} w_{k'} y_{k'}.$$

et $N_{i,sp}^{ps,WR1}$ a la même définition avec y_k fixé à 1. L'espérance de $\hat{y}_i^{ps,WR1}$ en fonction de la couverture, du degré de faible effectif (*sparseness*) et du plan d'échantillonnage est $E_{i,sp} E_{i,sp}(\hat{y}_i^{ps,WR1}) = T_c^g + (f_{\max}^g - 1) T_c^g$ où $T_c^g = \sum_{i \in A_{i,sp}^g} p_i T_c^i$. De même, $E_{i,sp} E_{i,sp}(\hat{y}_i^{ps,WR1}) = N_c^g + (f_{\max}^g - 1) N_c^g$ avec $N_c^g = \sum_{i \in A_{i,sp}^g} p_i N_c^i$. Après certaines manipulations, le biais approximatif de $\hat{y}_i^{ps,WR1}$ devient

$$(6) \quad \text{biais}(\hat{y}_i^{ps,WR1}) = \sum_{i \in A_{i,sp}^g} \sum_{h, j \in s_h} \sum_{k \in s_{hj(i)}} \frac{(\alpha \phi)^g}{C_{\alpha \phi, y, g}^g}.$$

où

$$(\alpha \phi)^g = \sum \alpha_i \phi_k / N^g, \alpha_i = 1 + (f_{\max}^g - 1) p_i, C_{\alpha \phi, y, g}^g = \sum (\alpha_i \phi_k / N^g) (y_k - \bar{y}^g) / N^g.$$

et les sommations sont faites sur $i \in A_{i,sp}^g, h, j \in U_h$, et $k \in U_{hj(i)}$. Dans le cas d'une probabilité de couverture commune dans la post-strate i , c'est-à-dire $\phi_k = \phi(i)$, nous

avons

$$(\alpha \phi)^g = \bar{\phi}^g + (f_{\max}^g - 1)(\bar{p})^g$$

et

$$\alpha_i \phi(i) - (\alpha \phi)^g = (\phi(i) - \bar{\phi}^g) + (f_{\max}^g - 1)(p_i \phi(i) - (\bar{p})^g)$$

où $(\bar{p})^g = \sum A_{i,sp}^g W_{i,sp}^g p_i \phi(i)$. Il découle de cela que

$$C_{\alpha \phi, y, g}^g = C_{\phi, y, g}^g + (f_{\max}^g - 1) C_{p, y, g}^g$$

avec

$$C_{p, y, g}^g = \sum A_{i,sp}^g W_{i,sp}^g p_i \phi(i) - (\bar{p})^g (y_i - \bar{y}^g).$$

Si les moyennes de cellules \bar{y}_i sont toutes égales dans un groupe, alors $C_{p, y, g}^g = 0$ et $\hat{y}_i^{ps,WR1}$ sera approximativement sans biais. Dans le cas particulier où la couverture est constante dans un groupe de cellules, c'est-à-dire $\phi(i) = \bar{\phi}^g$, alors $C_{\alpha \phi, y, g}^g = (f_{\max}^g - 1) \bar{\phi}^g \sum_{i \in A_{i,sp}^g} W_{i,sp}^g p_i$. Donc, si p_i et $(p_i - \bar{p}^g)(y_i - \bar{y}^g)$ sont constantes dans $A_{i,sp}^g$, alors $\hat{y}_i^{ps,WR1}$ sera presque sans biais, même si les moyennes de cellules $\bar{y}_i, i \in A_{i,sp}^g$

2.2 Moyenne poststratifiée sans regroupement de cellules

La moyenne poststratifiée est définie comme étant $\hat{y}^{ps1} = 1/N \sum_{i=1}^I (N_i^u/N_i^t) \bar{y}_i^u$ où \bar{y}_i^u et N_i^u sont définis comme dans (1), mais en excluant la sommation sur i . Définitions $T_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k y_k$ et $N_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k$. Il s'agit des valeurs prévues (par rapport au scénario de couverture) du total et du nombre d'unités couvertes dans la post-strate i . Si nous développons \hat{y}^{ps1} en série autour de (T_i^u, N_i^u) , $i = 1, \dots, I$, son approximation linéaire est

$$\hat{y}^{ps1} \doteq \frac{1}{N} \left[\sum_{i=1}^I \frac{N_i^t}{N_i^u} T_i^u + \sum_{i=1}^I \frac{N_i^t}{N_i^u} \left(\bar{y}_i^u - \frac{T_i^u}{N_i^u} N_i^u \right) \right].$$

Le biais de l'estimateur poststratifié est alors égal au premier terme de cette expression moins $\sum_{i=1}^I T_i^u/N$ et, après certaines manipulations, il peut s'écrire

$$\text{biais}(\hat{y}^{ps1}) \doteq \sum_{i=1}^I W_i^t \frac{\phi_i}{C^{\phi i t}} \quad (3)$$

où $\bar{\phi}_i = \sum \phi_k/N_i^t$, $C^{\phi i t} = \sum (\phi_k - \bar{\phi}_i)(y_k - \bar{y}_i^u)/N_i^t$, et \sum dénote la somme sur $h, j \in U_h$, et $k \in U_{h(j)}$. Donc, \hat{y}^{ps1} est biaisé en cas de toute corrélation entre la variable y mesurée et la probabilité de couverture ϕ_k dans n importe quelle post-strate. Si le taux de couverture est constant, à $\phi_k = \phi(i)$ dans la post-strate i , alors l'estimateur poststratifié est approximativement sans biais. De (3), il découle que les post-strates devraient être formées de façon que leurs taux de couverture ou bien les y soient homogènes dans chacune d'elles. Cette condition est comparable aux recommandations faites par Eltinge et Yansaneh (1997), Kalton et Malting (1991), ainsi que Little et Vartivarian (2005) pour la création des cellules pour la correction de la non-réponse. Dans les grandes enquêtes, l'ensemble initial de post-strates candidates est souvent plus étendu que ne le permet l'échantillon. Sauf rares exceptions, on regroupe certaines post-strates initiales afin de contrôler les corrections des poids. Habituellement, la raison pour laquelle aucun regroupement n'est effectué est que de petites catégories ont été combinées au préalable en s'appuyant sur l'expérience acquise antérieurement dans le cadre de la même enquête ou d'une enquête comparable. En ce sens, la poststratification sans regroupement de cellules, PS1, n'existe pas vraiment en pratique. La poststratification avec approche plus fréquente,

2.3 Moyenne poststratifiée avec regroupement de cellules

Examinons maintenant l'estimateur poststratifié avec regroupement de cellules, dans le cadre duquel les cellules à faible effectif sont repérées et combinées à d'autres cellules

- 1) calculer les critères de regroupement pour chaque cellule, c'est-à-dire le FCI, N_i^u/N_i^t , et la taille d'échantillon;
- 2) repérer les cellules à faible effectif, c'est-à-dire celles dont les critères tombent en dehors des limites pour le regroupement;
- 3) pour chaque cellule à effectif non faible, déterminer la voisine à effectif non faible la plus proche et combiner les deux cellules.

La moyenne poststratifiée avec regroupement de cellules est alors donnée par $\hat{y}^{ps2} = 1/(N \sum_{i=1}^I (N_i^u/N_i^t) \bar{y}_i^u)$ où $\bar{y}_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k y_k / \pi_k$ et $N_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k$. Définissons $T_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k y_k$ et $N_i^u = \sum_{h,j \in U_h, k \in U_{h(j)}} \phi_k$. Le développement en série de \hat{y}^{ps2} autour de (T_i^u, N_i^u) pour chaque groupe g donne

$$\hat{y}^{ps2} \doteq \frac{1}{N} \left[\sum_{i=1}^I \frac{N_i^t}{N_i^u} T_i^u + \sum_{i=1}^I \frac{N_i^t}{N_i^u} \left(\bar{y}_i^u - \frac{T_i^u}{N_i^u} N_i^u \right) \right].$$

Il s'ensuit que

$$\text{biais}(\hat{y}^{ps2}) \doteq \sum_{i=1}^I W_i^g \frac{\phi_i}{C^{\phi i g}} \quad (4)$$

avec

$$W_i^g = N_i^g/N_i^t, \bar{\phi}_i = \sum \phi_k/N_i^g, C^{\phi i g} = \sum (\phi_k - \bar{\phi}_i)(y_k - \bar{y}_i^g)/N_i^g$$

et les sommations on $\bar{\phi}_i$ et $C^{\phi i g}$ sont faites sur $i \in A_i^g, h, j \in U_h$, et $k \in U_{h(j)}$. Si ϕ_k est constant dans le groupe g , cet estimateur est sans biais, mais si $\phi_k = \phi(i)$, c'est-à-dire si le taux de couverture est constant dans la post-strate i , mais qu'il peut différer selon la post-strate, alors le biais devient

$$\text{biais}(\hat{y}^{ps2}) \doteq \sum_{i=1}^I W_i^g \frac{\phi_i}{C^{\phi i g}} \quad (5)$$

avec $\bar{\phi}_i = \sum W_i^g \phi(i)$, $C^{\phi i g} = \sum W_i^g (\phi(i) - \bar{\phi}_i)(\bar{y}_i^g - \bar{y}_i^g)$, $W_i^g = N_i^g/N_i^t$, et les sommations sont faites sur $i \in A_i^g$.

$\bar{Y} = \sum_{i=1}^I T_i / N$ où T_i est le total pour la population complète d'unités dans la post-strate i (et pas seulement la partie couverte). Après certains calculs, le biais est

$$(2) \quad \text{biais}(\hat{\phi}_\pi) \doteq \frac{N^c}{T^c} - \frac{1}{N} \sum_{i=1}^N T_i = \frac{\phi}{C^\phi}$$

où $\phi = \sum \phi_k / N$, $C^\phi = \sum (\phi_k - \phi)(Y_k - \bar{Y}) / N$, $\bar{Y} = T / N$, et \sum dénote la somme sur $i, h, j \in U_h$, et $k \in U_{h(i)}$. Par conséquent, \hat{Y}_π est biaisé s'il existe toute corrélation entre la variable mesurée, Y , et la probabilité de couverture, ϕ_k . Dans (2), le biais est $O(1)$, si bien qu'il reste important même dans les grands échantillons.

Si la probabilité de couverture est la même pour chaque unité de la post-strate i , c'est-à-dire $\phi_k = \phi(i)$ pour tout $k \in U_{h(i)}$, alors le biais approximatif se réduit à

$$\text{biais}(\hat{Y}_\pi) = \phi_{\pi}^{-1} \sum_i W'_i (\phi(i) - \phi) (X_i - \bar{X}) \quad \text{ou } W'_i = N_i / N \text{ et } \bar{X}_i = \sum_{h,j \in U_h, k \in U_{h(i)}} Y_k / N_i.$$

Si l'existence d'une corrélation entre les probabilités de couverture de post-strate et les moyennes de post-strate, l'estimateur de Hájek produira de nouveau un biais qui pourrait être positif ou négatif. Si les taux de couverture ou les moyennes ne varient pas selon la strate, c'est-à-dire $\phi(i) = \phi_o$ ou $\bar{X}_i = \bar{X}$, l'estimateur de Hájek sera sans biais, mais les post-strates ne sont habituellement pas formées de cette façon. En outre, le biais existe même si l'échantillonneur ne sait pas quel est l'ensemble approprié de post-strates qui subdivisent la population en groupes ayant des moyennes différentes.

dans la post-strate i ; U_h est la population d'UPB dans la strate h ; U_{hj} est la population d'unités dans l'UPB j dans la strate h qui sont présentes dans la post-strate i . Pour les besoins de l'analyse exposée à la présente section, il n'est pas nécessaire de spécifier le plan d'échantillonnage plus en détail. Notons qu'aux sections 2 et 3, certaines sommes dans la forme $\sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} \dots$ pour les unités comprises dans la post-strate i pourraient être simplifiées en notation plus élaborée afin de montrer clairement comment les degrés d'échantillonnage devraient être traités.

2.1 Estimateur de Hájek

En premier lieu, considérons l'estimateur de Hájek d'une moyenne, qui est

$$(1) \quad \hat{Y}_\pi = \frac{\left(\sum_{i=1}^I \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} Y_k / \pi_k \right)}{\left(\sum_{i=1}^I \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} 1 / \pi_k \right)} \doteq \bar{Y}_\pi / N_\pi.$$

L'espérance de \hat{Y}_π par rapport à l'échantillonnage et au scénario de couverture est $E_{\pi} E_{\pi}(\hat{Y}_\pi) = \sum_{i=1}^I \sum_h \sum_{j \in s_h} \phi_k Y_k \equiv T^c$, où l'indice supérieur c signifie « couvert ». De même, l'espérance de N_π est $E_{\pi} E_{\pi}(N_\pi) = \sum_{i=1}^I \sum_h \sum_{j \in s_h} \sum_{k \in s_{h(i)}} 1 \equiv N^c$. Si nous développons \hat{Y}_π autour de (T^c, N^c) , son approximation linéaire est $\hat{Y}_\pi \doteq T^c / N^c + 1 / N^c (T^c - T^c) / N^c$. Considérons maintenant le biais de \hat{Y}_π comme un estimateur de

Tableau 1
Ratios de couverture pour la Current Population Survey (CPS) et la Behavioral Risk Factors Surveillance Survey (BRFSS). Blanche uniquement et noire uniquement signifient que seules ces races ont été déclarées par un répondant à la CPS. Le groupe racial résiduel uniquement comprend les cas indiquant une seule race autre que blanche ou noire, ainsi que les cas déclarant deux races ou plus. Dans le tableau pour la BRFSS, les Hispaniques peuvent être de n'importe quelle race

Current Population Survey de mars 2002									
Âge		Blanche uniquement		Noire uniquement		Résiduel uniquement			
Hommes		Femmes		Hommes		Femmes		Hommes	
0 à 15 ans	0,93	0,93	0,78	0,79	0,91	0,93	0,75	0,80	0,84
16 à 19 ans	0,90	0,88	0,76	0,81	0,93	0,75	0,72	0,76	0,80
20 à 24 ans	0,79	0,85	0,72	0,77	0,75	0,72	0,76	0,80	0,84
25 à 34 ans	0,83	0,89	0,70	0,76	0,80	0,76	0,80	0,84	0,88
35 à 44 ans	0,28	0,31	0,37	0,25	0,18	0,30	0,39	0,22	0,31
45 à 54 ans	0,19	0,26	0,12	0,24	0,15	0,22	0,39	0,22	0,31
55 à 64 ans	0,20	0,31	0,10	0,16	0,19	0,39	0,22	0,31	0,31
65 à 74 ans	0,25	0,31	0,25	0,20	0,18	0,30	0,39	0,22	0,31
Tous âges confondus	0,25	0,31	0,25	0,20	0,18	0,30	0,39	0,22	0,31
Sources : Bureau of the Census (2002), Gonzalez, Town et Kim (2005).									

Regroupement de cellules lors de la poststratification

Jay J. Kim, Jianshu Li et Richard Valliant¹

Résumé

La poststratification est une méthode courante d'estimation dans le cas des enquêtes-ménages. Les cellules sont créées d'après les caractéristiques qui sont connues pour tous les répondants de l'échantillon et pour lesquelles il existe des dénombrements de contrôle externes provenant d'un recensement ou d'une autre source. La couverture de certains groupes démographiques peut poststratification sont habituellement appelés ratios de couverture. La couverture de certains groupes démographiques peut être sensiblement inférieure à 100 % et la poststratification est destinée à corriger les biais résultant d'une couverture inadéquate. Une méthode standard de poststratification consiste à regrouper ou à combiner certaines cellules lorsque les tailles d'échantillon sont inférieures à un minimum donné ou que les corrections des poids sont supérieures à un maximum donné. Le regroupement peut accroître ou réduire la variance d'une estimation, mais pourrait simultanément augmenter son biais. Nous étudions les effets, sur le biais et la variance, de ce type de regroupement dynamique des cellules du point de vue théorique et par simulation en utilisant une population basée sur la National Health Interview Survey de 2003. Nous proposons deux estimateurs possibles qui réduisent l'importance des corrections des poids lorsque les cellules sont regroupées.

Mots clés : Biais; combinaison de cellules; erreur de couverture; poststratification; sous-couverture; élagage des poids.

1. Introduction

La poststratification est une technique fréquemment adoptée dans la pondération des enquêtes afin 1) de réduire les variances ou 2) de corriger la couverture insuffisante de l'échantillon pour certains groupes de population cibles. Dans le cas des enquêtes-ménages aux États-Unis, le deuxième objectif est particulièrement important, parce que certains groupes démographiques, comme les jeunes hommes noirs, sont moins bien couverts que les autres (par exemple, voir Kostanich et Dippo 2000, chapitre 16). La correction pour la sous-couverture peut aboutir à des poids variables qui corrigent peut-être le biais, mais accroissent aussi les erreurs-types. Souvent, les praticiens évitent de procéder à des corrections extrêmes des poids, sacrifiant effectivement une certaine réduction du biais afin de maîtriser les variances.

Une façon de contrôler la grandeur des corrections des poids consiste à regrouper les cellules résultant de la poststratification initiale si la correction dans l'une d'elles est supérieure à une limite donnée. Little (1993), ainsi que Lazzeroni et Little (1998) s'intéressent aux méthodes de regroupement des catégories des variables de post-stratification ordinaires. D'autres stratégies de regroupement des strates ou de construction d'estimateurs ont été proposées par Fuller (1986), Kalton et Maligaglia (1991), Vajs (1982) décrivent certaines applications pratiques. Dans le présent article, nous étudions les effets de la combinaison des cellules sur le biais et la variance, en présupposant que des

cellules définies à un niveau plus fin de détail seraient préférables si les tailles d'échantillon et les corrections des poids se situaient dans certaines limites établies par les concepteurs de l'enquête.

Deux critères sont souvent utilisés pour décider si une cellule doit être ou non regroupée avec une autre. Le premier est le *ratio de couverture inverse* ou *facteur de correction initial* (FCI), qui est défini comme étant le ratio du dénombrement de contrôle au dénombrement d'échantillon initialement pondéré pour la cellule. Un ratio significativement différent de 1 indique que la couverture est trop faible ou trop élevée pour le groupe représenté par la cellule. Lorsque le FCI d'une cellule se situe en dehors de limites établies d'avance, on la combine à une autre. Par exemple, le seuil de regroupement pour un ratio « élevé » pourrait être égal à 2 et celui pour un ratio « faible » à 0,6, valeurs qui sont les seuils utilisés pour la Current Population Survey (CPS) réalisée par le Bureau of the Census des États-Unis (voir Kostanich et Dippo 2000, page 10-7). Le deuxième critère est la taille d'échantillon. Une cellule dont le dénombrement brut d'échantillon est trop faible peut être regroupée avec une autre en prenant pour motif que le FCI est instable. Nous dirons qu'une cellule possède un *faible effectif* si elle viole l'un des deux critères susmentionnés et est groupée avec une autre.

Les catégories de variables qui définissent les post-strates sont habituellement triées suivant un ordre naturel (par exemple, les catégories d'âge ou de revenu) ou un ordre commode (par exemple, race-appartenance ethnique). En pratique, on regroupe généralement une cellule avec une

1. Jay J. Kim, National Center for Health Statistics, Centers for Disease Control and Prevention, Jianshu Li, Joint Program in Survey Methodology, University of Maryland, Richard Valliant, Survey Research Center, University of Michigan.

Durant la phase d'initialisation, il faut adapter les pondérations. En 2005, les individus de $\Omega_{2004,2005}^{(1)mnig}$ auront un poids final transversal directement issu du tirage du logement dans $a_{2005,1}^{(1)}$ (ils ne peuvent être atteints qu'au travers de ce panel entrant). En revanche, tous les autres individus sont « normalement » enquêtés à partir des neufs panels $a_{2005,k}^{(1)}$ ($1 \leq k \leq 9$), si bien que leurs poids issus du partage des poids seront tous systématiquement divisés par 9. En 2006, les individus de $\Omega_{2005,2006}^{(1)mnig}$ auront un poids égal à celui du logement dans lequel ils résident et qui reflète directement le tirage de $a_{2006,1}^{(1)}$ ceux de $\Omega_{2004,2006}^{(1)mnig}$ auront leurs poids issus du partage des poids divisés par 2, et tous les autres individus auront leurs poids issus du partage des poids divisés par 9.

Ce traitement s'effectue bien sous-échantillon par sous-échantillon et ne doit pas tenir compte de ce qui survient dans les autres sous-échantillons. Si un individu est enquêté à t par l'intermédiaire de deux (ou plus) sous-échantillons $a_{t,k}$ distincts, on détruit le traitement complet associé à chacun des deux (ou plus) sous-échantillons. Ce peut être le cas, par exemple, d'un ménage composé de deux individus-panels provenant de deux sous-échantillons $a_{t,k}$ différents parce que ces individus se sont mariés et qu'avant leur mariage ils étaient suivis chacun séparément en formant un ménage de taille un. Dans cette configuration, chacun des deux individus est « formellement » enquêté deux fois, une fois en tant qu'individu-panel, une fois en tant que cohabitant.

Finalement, pour l'estimation de la différence $\Delta_{t,t+1}^{(1)}$ $Y_{t+1}^{(1)} - Y_t^{(1)}$, on pourra utiliser les poids $W_{t,t+1}^{(1)}$ issus de la méthode 1, et ainsi calculer

$$\Delta_{t,t+1}^{*} = \sum_{i \in n_{t+1}} W_{t+1,(1)}^{(1)} Y_{t+1}^{(1)} - \sum_{i \in n_t} W_{t,(1)}^{(1)} Y_t^{(1)} \quad (13)$$

Si on, on pourra utiliser les poids $W_{t,t+1}^{(2)}$ issus de la méthode 2. L'estimateur de la différence $\Delta_{t,t+1}^{*}$ sera alors donné par

$$\Delta_{t,t+1}^{*} = \sum_{i \in n_{t+1}} W_{t+1,(2)}^{(2)} Y_{t+1}^{(2)} - \sum_{i \in n_t} W_{t,(2)}^{(2)} Y_t^{(2)} \quad (14)$$

Bibliographie

Ardilly, P. (2006). *Les techniques de sondage*, 2^{ème} édition. Éditions Technip, Paris.

Lavallée, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

Lavallée, P. (2002). *Le sondage indirect, ou la méthode généralisée du partage des poids*. Éditions de l'Université de Bruxelles et Éditions Ellipses.

Lévesque, I. et Franklin, S. (2000). Longitudinal and Cross-Sectional Reference year. Document de recherche sur le revenu, Statistique Canada, Catalogue No. 75F0002MIE-00004, juin 2000.

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprès des ménages à panels multiples. *Techniques d'enquêtes*, 27, 189-200.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, M.P., Drew, J.D., Gambino, J.G. et Mayda, F. (1990). *Méthodologie de l'Enquête sur la population active*. Statistique Canada, Catalogue 71-526.

échantillonnable à $t-8$, il en sera de même à toute date comprises entre $t-8$ et $t-1$. Cela revient à négliger les situations où un individu du champ à une date donnée en sort durant quelque temps (émigration, par exemple), puis y revient ensuite.

Par ailleurs, on note $u_{t,k}^i$ l'échantillon transversal à t issu du panel $a_{t,k}^i$, ce qui conduit à $\bigcup_{k=1}^9 u_{t,k}^i = u_t^i$. Soit $\Omega_{a,t,t}^i$ le total des $Y_{t,t}^i$ défini sur $\Omega_{a,t,t}^i$. On a alors, suite au

partage des poids effectué pour tout $k = 2, \dots, 9$:

$$E \left(\sum_{j \in u_{t,k}^i} W_j^i(t, k) \cdot Y_{t,t}^i \right) = \sum_{j \in \Omega_{a,t,t}^i} Y_{t,t}^i - \sum_{l=1}^{\alpha-1} Y_{a,t,t}^i \quad (9)$$

et

$$E \left(\sum_{j \in u_{t,1}^i} W_j^i(t, 1) \cdot Y_{t,t}^i \right) = \sum_{j \in \Omega_{a,t,t}^i} Y_{t,t}^i = Y_t^i \quad (10)$$

puisque $u_{t,1}^i = a_{t,t}^i$.

Avec un système de panels à courte durée, on pourrait peut-être négliger les $\Omega_{a,t,t}^i$ devant le vrai total sur $\Omega_{t,t}^i$ et

serait $W_j^i(t, k)/9$ si t est issu de $a_{t,k}^i$, ce qui conduirait à l'estimateur final

$$Y_t^i = \frac{1}{9} \sum_{k=1}^9 \sum_{j \in u_{t,k}^i} W_j^i(t, k) \cdot Y_{t,t}^i \quad (11)$$

Les panels utilisés en France ont cependant une durée de vie longue, aussi il est fort possible que l'on ne puisse pas raisonner ainsi (l'examen des fichiers de collecte permettra d'en juger) et qu'il soit nécessaire de pondérer spécifiquement les individus des $\Omega_{a,t,t}^i$. Dans ces conditions, on vérifie que tout individu i de $\Omega_{a,t,t}^i$ qui se trouve finalement dans l'échantillon transversal u_t^i aura un poids transversal brut $W_{t,t}^{i(2)}$ égal à la valeur $W_{t,t}^i(k)$ issue du partage des poids, divisée par $t-\alpha$ (et donc $1 \leq t-\alpha \leq 8$). Pour sa part, tout individu de $\Omega_{t,t}^i$ qui n'appartient à aucun des $\Omega_{a,t,t}^i$ (donc la grande majorité des cas) aura un poids final égal à $W_{t,t}^i(k)/9$. On remarquera par ailleurs que si i se trouve dans $\Omega_{a,t,t}^i$, il ne peut être enquêté qu'au travers de $a_{t,1,t}^i, a_{t,2,t}^i, \dots, a_{t,t-\alpha,t}^i$. On obtient ainsi

$$W_{t,t}^{i(2)} = \begin{cases} W_{t,t}^i(k)/(t-\alpha) & \text{si } i \in \Omega_{a,t,t}^i \\ \text{sinon} & \end{cases} \quad (12)$$

facilement, mais qui se heurte à une difficulté qui n'apparaissait pas dans la méthode précédente et qui risque en pratique de rendre la pondération définitive un peu moins rigoureuse. L'idée est de raisonner non pas sur l'ensemble des sous-échantillons, mais sous-échantillon par sous-échantillon. On considère un quelconque des neuf sous-échantillons $a_{t,k}^i$ ainsi que l'échantillon de ménages auquel il mène. On applique alors le partage des poids, ce qui donne en régime stationnaire une pondération individuelle égale à

$$W_j^i(t, k) = \frac{\sum_{j \in m} W_j^i(t, k)}{\sum_{j \in a_{t,k}^i} 1} \quad (8)$$

pour tout individu i du ménage m . On vérifie très facilement que si $k = 1$ (cas du sous-échantillon entrant), $W_j^i(t, 1)$ est le

La difficulté associée à cette approche est liée à l'existence (a priori) à la date t d'individus non enquêtés

parce qu'ils appartiennent à des ménages qui ne sont pas du tout « atteignables » au travers de l'échantillonnage $a_{t,k}^i$

(des lors que $k \geq 2$), c'est-à-dire qui ont une probabilité nulle d'être enquêtés à t . Ce phénomène perturbeait

n'existant pas dans la méthode précédente grâce à la prise en compte globale de l'ensemble des sous-échantillons

puisque à la date t tout ménage a une probabilité strictement positive d'être sélectionné, au moins au travers de $a_{t,1}^i$.

C'est une nouvelle occasion de souligner un des atouts essentiels de l'échantillonnage rotatif qui constitue une

technique permettant chaque année de couvrir l'intégralité de la population. Dans notre approche, il est clair que si on considère $a_{t,k}^i$ pour $k \geq 2$, on ne couvre pas la population

des ménages constitués exclusivement d'« immigrants » (au sens large) et $t-k+1$ et t . Pour formaliser le contexte et

aboutir au poids transversal final, on notera $\Omega_{a,t,t}^i$ la population d'« immigrants » (au sens large) présente à t

dans des ménages ne comprenant que des immigrants échantillonnables après l'année α , avec $t-8 \leq \alpha \leq t-1$.

Notons que, plus précisément, il faudrait dire « échantillonnables à partir d'une date strictement postérieure à la

date de collecte de l'année α ».

A la date t , la population complète Ω_t^i est partitionnée en neuf composantes : les huit sous-populations $\Omega_{a,t,t}^i$, avec α variant de $t-8$ à $t-1$, et la sous-population constituée par les individus, soit qui étaient déjà enquêtés à $t-8$, soit qui sont devenus enquêtés à une date ultérieure à $t-8$ (donc des immigrants au-delà de $t-8$) mais qui sont intégrés à t dans un ménage comprenant au moins une personne enquêtée à $t-8$. Notons que l'on considère que si le ménage à t comprend au moins une personne

Comme dans le cas longitudinal (voir 3.1.), la pondération ne peut s'effectuer que si le système informatique de gestion

panel de u_i^k à l'ensemble des échantillons panels a_i^k dans lesquels il se trouve. Au dénominateur, on dénombre pour chacune des neuf années $t = 8 - 4$ à t considérées, les individus du ménage (qu'ils soient individus-panels ou cohabitants) que se trouvent dans la base de sondage utilisée pour le tirage du sous-échantillon panel entrant l'année en question. Ce calcul nécessite évidemment la disponibilité de l'information via le questionnaire.

Cette approche a un double atout : d'une part elle est parfaitement générale, et d'autre part elle donne immédiatement lieu à des poids transversaux sans biais parce que tout ménage transversal est nécessairement relié à l'un quelconque des neuf sous-échantillons considérés. Le fait qu'il y ait chaque année un sous-échantillon entrant permet de représenter l'intégralité de la population transversale Ω_{2004} , c'est-à-dire, dans un langage plus technique, assure l'existence d'au moins un lien pour chaque ménage considéré à t . C'est une propriété intéressante de l'échantillonage rotatif que nous avons déjà mentionnée à la section 2.2. En contrepartie, la formule de pondération a un inconvénient qui est sa (relative) complexité, à la fois sur le plan théorique et lors de la phase de programmation informatique.

Dans la phase d'initialisation (donc jusqu'en 2011 compris), cette expression doit être adaptée : le numérateur ne change pas mais le dénominateur dénombre les individus échantillonnables à partir de 2004, première année de réalisation de l'enquête. En 2004, la pondération est évidente puisqu'il n'y a pas de partage des poids, mais en 2005 on prendra

$$(6) \quad W_{f_j(t,k)}^{(t)} = \frac{\sum_{k=1}^K \sum_{j \in \Omega_{2004}} W_j(t,k)}{\sum_{k=1}^K \sum_{j \in \Omega_{2005}} 1 + 8 \cdot \sum_{j \in \Omega_{2004}} 1}.$$

En 2006, ce sera

$$(7) \quad W_{f_j(t,k)}^{(t)} = \frac{\sum_{k=1}^K \sum_{j \in \Omega_{2006}} 1 + \sum_{k=1}^K \sum_{j \in \Omega_{2005}} 1 + 7 \cdot \sum_{j \in \Omega_{2004}} 1}{\sum_{k=1}^K \sum_{j \in \Omega_{2006}} 1 + \sum_{k=1}^K \sum_{j \in \Omega_{2005}} 1 + 7 \cdot \sum_{j \in \Omega_{2004}} 1}.$$

4.2 Méthode 2

On peut avoir une vision alternative de la pondération transversale qui conduit à une expression de poids « un peu » plus simple et qui peut se programmer plus

du statut de « naissance » se fait généralement en demandant, pour les nouveaux-nés, la date de naissance, et pour les immigrants, la date d'entrée au pays. On ajoute qu'en pratique, le défaut de couverture des naissances est en général considéré comme négligeable parce qu'il est en partie corrigé par l'utilisation du redressement. La technique centrale utilisée pour produire les poids transversaux est la méthode de partage des poids (Lavallée 2002). Rappelons qu'à l'année t , on dispose de neuf sous-échantillons panels a_i^k ($1 \leq k \leq 9$). On présente ici deux approches possibles pour l'application de la méthode du partage des poids. Notons que les informations à recueillir dans le questionnaire sont identiques pour mettre en oeuvre les deux méthodes.

4.1 Méthode 1

L'approche la plus rigoureuse consiste à relier l'ensemble des neuf sous-échantillons a_i^k à l'échantillon transversal de l'année t , que nous noterons u_i^t (Merkourts 2001). L'échantillon u_i^t correspond donc à $s_i^t = \bigcup_{k=1}^9 a_i^k$. Il faut tout d'abord commencer par définir les liens associés à ce schéma : lorsqu'un individu-panel quelconque de l'un des neuf sous-échantillons a_i^k a été désigné par le sort, il pointe sur lui-même en tant qu'individu de l'échantillon transversal à t (schéma voisin de celui du 3.1.). Dans ces conditions et en régime stationnaire, le poids transversal $W_{f_j(t,k)}^{(t)}$ d'un individu quelconque i de u_i^t s'obtient de la façon qui suit. On note m le ménage auquel appartient i . On a

$$(5) \quad W_{f_j(t,k)}^{(t)} = \frac{\sum_{k=1}^K \sum_{j \in \Omega_{2004}} W_j(t,k)}{\sum_{k=1}^K \sum_{j \in \Omega_{2004}} 1}.$$

où $W_{f_j(t,k)}^{(t)}$ est le poids de sondage qui découle de l'échantillonage a_i^k . Cette expression montre que tous les individus d'un même ménage ont à la fin le même poids. Au numérateur, reflètent l'échantillonage) de tous les individus-panels du ménage, étant entendu qu'en général un individu-panel n'apparaît que dans un seul sous-échantillon mais qu'il peut y avoir des cas où un individu-panel a été tiré deux fois ou même davantage sur une période de neuf années consécutives (pour cause de démenagement, essentiellement). Il est à noter que l'échantillonage de logements dans l'échantillon-matériau et la BSLN s'appuie sur un principe de non-réinterrogation des logements déjà tirés et ainsi, dans le cas

de ménage apparaissant dans deux panels distincts est nul.

in fine cette expression paraît tout à fait opportun. Si on se place dans un cadre idéal - qui paraît néanmoins trop simplifié dans notre contexte - où la population n'évolue pas dans le temps, on aura $L_i = 8$ pour tout i . Notons cependant que la population évolue beaucoup en neuf années, mais avec des durées de panelisation plus courtes, ce cas idéal peut être une approximation acceptable. Si, de plus, les panels sont tirés à probabilités égales, W_i sera égal à une constante W et alors

$$(3) \quad W_{i,t+1}' = \frac{8}{W}.$$

Ce cas de figure reste très peu probable dans le cas de la France. D'une part, jusqu'en 2012, il y a coexistence de

sous-échantillons tirés avec des poids bruts nettement distincts (voir l'introduction). D'autre part, on aura tendance à concevoir l'échantillonnage en fixant le nombre total de logements à tirer (alors même que le nombre total de logements augmente) et non pas en raisonnant sur un objectif de taux de sondage constant.

Notons que la formule (3) est intuitive : finalement, tout se passe « comme si » n'importe quel individu de l'échantillon longitudinal $s_{t,t+1}'$ avait une probabilité de sélection égale à huit fois celle qui caractérise chaque sous-

Ce qui précède s'applique au régime stationnaire et doit être légèrement adapté durant la phase d'initialisation du processus, c'est-à-dire jusqu'en 2012. La première opération de nature longitudinale porte sur les données conjointes 2004-2005, pour estimer des évolutions entre 2004 et 2005 avec la population de référence 2004 (privée des « morts » en 2005). Dans ce contexte, il suffit de diviser tous les poids W_i des huit sous-échantillons $a_{2004,1}^{2005,1}$ par huit - autrement dit $L_i = 8$ pour tout i . En 2006, lorsqu'on s'intéressera aux évolutions 2005-2006, le dénominateur L_i pourra prendre deux valeurs. Dans le premier scénario, l'individu-panel i était dans la base de sondage utilisée en 2004 (donc potentiellement tirable en 2004) et alors $L_i = 8$. Ceci vient du fait que tout se passe comme si, en 2004, on avait effectué les sept tirages des panels $a_{2005,2}^{2005,2}$ à $a_{2005,8}^{2005,8}$ exactement dans les mêmes conditions. Dans le second scénario, l'individu i n'était pas dans la base de sondage de 2004 - mais alors il est dans la base 2005 et il se trouve nécessairement dans $a_{2005,1}^{2005,1}$ - pour être égal à 1, 2 ou 8, et ainsi de suite. Pour retrouver l'ensemble des valeurs possibles de L_i parmi $\{1, 2, 3, \dots, 8\}$, il faudra attendre la mesure des évolutions 2011-2012.

Passée cette étape de pondération longitudinale, on obtient des poids longitudinaux $W_{i,t+1}'$ et on forme l'estimateur de la différence $\Delta_{i,t+1}'$ selon

$$(4) \quad \Delta_{i,t+1}' = \sum_{t_g=t+1} W_{i,t+1}' \cdot (Y_{i,t+1}' - Y_i').$$

A priori, les poids $W_{i,t+1}'$ ne sont utilisés que dans le cadre d'une estimation d'évolution. Pour des estimations ponctuelles, ils apparaissent sans intérêt parce que la population d'inférence n'a pas grande signification à date donnée. Rappelons que jusqu'ici les $W_{i,t+1}'$ n'ont fait l'objet d'aucune correction pour non-réponse, ni redressement. En pratique, l'estimateur (4) pour l'enquête SILC sera sujet à de tels ajustements.

L'estimation de la différence $\Delta_{i,t+1}' = Y_{i,t+1}' - Y_i'$ correspond à une vision transversale : elle fait ainsi appel à la pondération qui est présentée dans la section suivante.

4. La pondération transversale

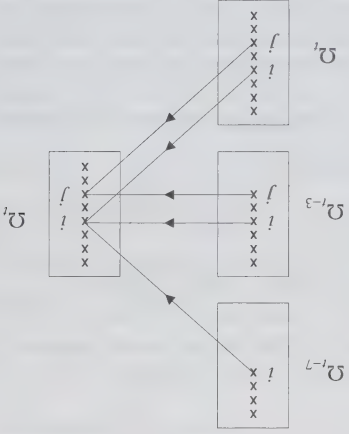
Il s'agit de pratiquer une inférence sur la population globale Ω_2 du champ de l'enquête à la date courante t . La difficulté essentielle tient au fait qu'un sous-échantillon (panelisé) donné ne couvre correctement, en théorie, la population que l'année de son tirage. Passée cette année, le sous-échantillon panel ne représente plus la population nouvelle des « naissances » c'est-à-dire ceux qui entrent dans le champ de l'enquête. Cela concerne en particulier les nouveaux-nés, les immigrants, les individus dont l'âge atteint certains seuils, les personnes anciennement sans domicile qui retrouvent un logement ordinaire, les retours de communautés, etc. Si en pratique on peut imaginer s'en saisir à posteriori pendant quelque temps, ce défaut de couverture devient assez vite excessif (cela est vrai chaque année pour la plupart des sous-échantillons panels) et il faut d'une façon ou d'une autre obtenir un échantillon complémentaire au panel. Il est à noter que la problématique de l'évolution dans le temps de la population est fortement dissymétrique parce que la sous-population qui disparaît d'une année sur l'autre (les « morts ») ne pose pas de problème particulier en terme de pondération.

Dans l'enquête SILC, l'échantillon complémentaire est obtenu en appliquant la méthodologie suivante : on décide, pour chaque individu-panel enquêté lors du processus de suivi longitudinal, d'interroger l'ensemble des individus du ménage dans lequel se trouve l'individu-panel. Ainsi, tout ménage enquêté dans l'optique transversale est composé de deux types de personnes : des individus panel et des cohabitants (ou nomme ainsi toute personne enquêtée qui n'est pas individu-panel). Cette méthodologie permet de couvrir une grande partie des « naissances » (au sens large) au sein de la population. Cependant, elle ne permet pas d'atteindre les ménages constitués seulement de « naissances » comme, par exemple, les ménages contenant seulement des immigrants. Précisons que la détermination

Ω_{t-k+1} . Pour SILC en France, il s'agit d'un ordre de grandeur de 60 millions. Le poids longitudinal à affecter à tout individu i de $s_{t,t+1}$ sera *in fine* :

$$(1) \quad W_{t,t+1}^i = \frac{1}{\sum_{k \in K_i} L_i} W_i(t, k).$$

Cette expression découle de l'application de la méthode de partage des poids (voir Lavallée 1995, ainsi que Lavallée 2002) où on définit la population initiale (celle des unités d'échantillonnage) comme réunion des populations $\Omega_{t-7}, \dots, \Omega_{t-1}, \Omega_t$ et la population finale (celle des unités d'observation) comme Ω_t . Le schéma ci-dessous illustre le contexte, où, pour plus de clarté, nous n'avons pas reproduit les huit sous-populations initiales, mais seulement trois d'entre-elles. Le nombre de liens apparaît alors clairement égal à L_i (ici, par exemple, i a exactement huit liens, mais j en a strictement moins de huit parce qu'il n'apparaît pas dans les bases de sondage les plus anciennes). Pratiquement, il est réaliste de faire comme si on avait $\Omega_{t-7} \subset \Omega_{t-6} \subset \dots \subset \Omega_{t-1} \subset \Omega_t$. On peut raisonner sur des individus qui emboîtent parce que, sauf exception, les individus qui sortent du champ au cours du temps avant t ne seront pas présents dans $s_{t,t+1}$.

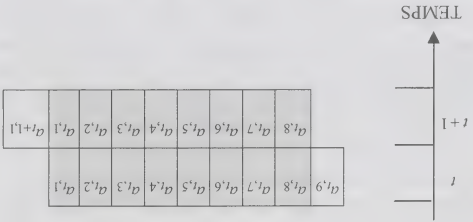


La formule (1) fournit l'expression la plus générale possible du poids longitudinal « brut ». On peut ensuite la simplifier dans différents contextes. Si par exemple on néglige les cas où un individu-panel peut être tiré deux fois ou plus, on a

$$(2) \quad W_{t,t+1}^i = \frac{L_i}{W_i}$$

où W_i est le poids de i relatif à l'unique sous-échantillon panel dans lequel il figure à la date t . Dans le cas de la France, compte tenu des tailles d'échantillon en jeu, adopter

On notera qu'on peut écrire $a_{t,k+1,k+1}^i = a_{t,k}^i$ ($\forall t, \forall k \neq 9$) puisque par principe on reprend intégralement chaque sous-échantillon panel (non sortant) d'une année sur l'autre. Schématiquement, on a :



La partie grisee représente $s_{t,t+1}^i$ qui est l'échantillon exploité dans cette approche longitudinale. C'est en effet sur les individus de $s_{t,t+1}^i$ que l'on peut obtenir à la fois les informations Y_t^i et Y_{t+1}^i sur l'individu i définies respectivement aux dates t et $t+1$.

Soit un individu i quelconque de Ω_t , dans le champ de l'enquête à t . On note L_i le nombre d'années parmi $\{t-7, t-6, \dots, t-1, t\}$ durant lesquelles l'individu i se trouvait dans le champ de l'enquête, donc était susceptible d'être tiré dans un panel « entrant ». Notons que l'on suppose ici que chaque année la base de sondage couvre exactement le champ de l'enquête. On a $L_i \in \{1, 2, 3, \dots, 8\}$. Par ailleurs, on note K_i l'ensemble des indices k parmi $1, 2, 3, \dots, 8$ pour lesquels on a $i \in a_{t,k}^i$. Il s'agit donc, à la date t , pour tout i de $s_{t,t+1}^i$, K_i sera par construction un ensemble contenant au moins un élément. La plupart du temps, K_i ne comprendra en fait qu'un seul indice, mais parfois il pourra en comprendre deux, voire davantage : en effet, ce cas surviendra si i est tiré dans un panel, qu'il déménage et que son nouveau logement est échantillonné dans un autre panel, une année ultérieure. Il est à noter que notre contexte exclut qu'un logement donné soit tiré deux fois, parce qu'il y a un principe de non-réinterrogation des logements de l'échantillon-maître et de la BS LN. Mais ce n'est qu'une convention de nature pratique, la théorie s'accommodant fort bien d'un système dans lequel on pourrait retirer les logements.

Si $i \in a_{t,k}^i$, appelons $W_i(t, k)$ son poids de sondage « brut » : il s'agit du poids de sondage du logement dans lequel se trouve i à la date de son tirage en tant qu'individu panel, donc lors du tirage annuel dans Ω_{t-k+1} . Ce système de poids permet une inférence directe du sous-échantillon $a_{t,k}^i$ vers la population complète Ω_{t-k+1} . En particulier, $\sum_{i \in a_{t,k}^i} W_i(t, k)$ estime sans biais le nombre total d'individus appartenant au champ de l'enquête et à la population

2.2 Enquêtes répétées dans le temps et stratégies

L'objectif est évidemment de pouvoir produire à la fois des estimations longitudinales et des estimations transversales. On peut envisager essentiellement trois stratégies :

- i. Un échantillonnage « indépendant » chaque année. En fait, compte tenu de l'existence d'un échantillon-mère et d'une base de sondage de logements neufs (BSLN), les tirages s'effectuent tous les ans dans les mêmes communes, et par conséquent il n'y a pas de véritable indépendance entre les différents sous-échantillons. Cette solution est largement perfectible en terme de précision des évolutions.
- 2. Un échantillonnage intégralement panelisé, c'est-à-dire un tirage initial d'échantillon interrogé chaque année. Ce scénario pose en particulier un problème de charge, car l'opération SILC est engagée pour une durée indéterminée. De ce fait, il est irréaliste.
- 3. Un échantillon rotatif. C'est ce scénario qui a été choisi, compte tenu des avantages qu'il présente pour satisfaire les attentes en matière à la fois longitudinale et transversale.

Le tableau qui suit qualifie les trois plans de sondage envisageables en fonction des deux approches souhaitées.

TYPE d'échantillon		Approche	
« Indépendant » chaque année	Panel	TRANSVERSALE	LONGITUDINALE
		NATUREL	POSSIBLE mais moins efficace
Rotatif	IMPOSSIBLE sans tirage complémentaire	POSSIBLE	POSSIBLE

La stratégie rotative présente quatre grands atouts :

- i. Elle réduit l'erreur d'échantillonnage associée à la mesure des évolutions (sur le principe, comme pour les panels, même si elle est moins efficace en théorie que le panel « pur »).
- ii. Elle limite la charge des enquêtes par rapport au panel « pur ». En la circonstance, s'agissant pour la France d'un panel de neuf années, cet argument doit être utilisé avec modestie. Il a cependant plus de force dans le scénario préconisé par Eurostat, qui donne lieu à une enquête annuelle durant quatre années consécutives.
- iii. Elle permet de prendre en compte d'une manière très « naturelle » l'évolution de la population avec le temps. Ce point sera plus compréhensible

3. La pondération longitudinale

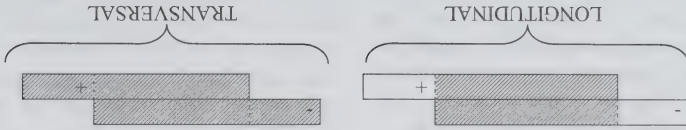
- i. Elle nécessite un suivi des individus dans le temps, ce qui occasionne des coûts de dépistage et des non-réponses du fait des déménagements.
- ii. Par nature, la longueur des séries individuelles se limite à neuf années, ce qui est déjà fort appréciable, mais évidemment moins riche qu'un pur panel.
- iii. La technique de pondération longitudinale/transversale n'est pas simple...

- iv. Elle permet de réduire les erreurs d'observation (comme les panels).
- lorsqu'on abordera la question de la couverture des populations nouvelles.

C'est une pondération a priori un peu plus simple à concevoir que la pondération transversale parce qu'il n'y a pas à tenir compte de l'évolution de la population dans le temps (en dehors des « morts », qui par convention disparaissent du champ avec le temps, mais ce point ne pose pas vraiment de problème technique). On rappelle que le principe de l'estimation longitudinale consiste à pratiquer une inférence sur la base d'une unique population considérée à une date initiale.

C'est clairement le caractère rotatif du plan de sondage qui complique la pondération, puisque entre deux années consécutives t et $t + 1$, on va mobiliser huit panels distincts, tirés dans des populations physiques qui sont, bien entendu, différentes d'une année sur l'autre). Si on ne manipulait qu'un seul panel, il suffirait de s'en tenir à l'utilisation des poids de sondage associés aux individus du panel encore dans le champ à la date t , ni plus ni moins, puisque ces poids sont calculés une fois pour toutes au moment du tirage et permettent chaque année, sur toute la durée de vie du panel, une inférence sur la population initiale.

La difficulté essentielle consiste à représenter la population Ω_t à la date t à partir de huit sous-échantillons panels tirés à des dates différentes. On peut comprendre intuitivement qu'un individu physique donné ait *in fine* une probabilité de sélection à la date t qui dépend du nombre de sous-échantillons panels dans lesquels il est susceptible d'être tiré. On suppose dans cette partie qu'il n'y a pas de non-réponse. La situation peut être formalisée de la manière suivante, en notant :



que Lèvesque et Franklin 2000).

Il est à noter que l'on n'abordera pas les questions de correction de la non-réponse ni du redressement des estimations. Le traitement de ces questions renvoie à ce que l'on retrouve en général dans les enquêtes longitudinales comme, par exemple, l'EDTR (voir Lavalée 1995, ainsi

Comme SILC est une enquête longitudinale où il y a chevauchement de panels dans le temps, la pondération de l'échantillon amène une problématique particulière. Cet article présente en détail les deux types de pondération utilisés pour SILC. On discutera en premier lieu de principes généraux reliés au plan de sondage de SILC. En deuxième lieu, on présentera la pondération longitudinale et

Le protocole de collecte permet de considérer chaque sous-échantillon comme un véritable panel d'individus : en effet, on suit physiquement les personnes qui quittent leur logement, les différentes directions régionales de l'INSEE se transmettant les dossiers des individus du panel qui démentagent. Pour plus de détails sur le plan de sondage de SILC, on peut consulter le *Journal Officiel de l'Union Européenne* du 17 novembre 2003 et les documents internes

La procédure d'échantillonnage proprement dite est la procédure standard utilisée lorsqu'on tire dans l'échantillon-précédent et dans la BSLN (voir Ardilly 2006). Dans le cas présent, il n'y a aucune surreprésentation de catégories d'individus. C'est une enquête à taux uniforme - aux arrondis près - à l'exception des logements vacants ruraux et des résidences secondaires au RP de 1999 qui sont devenus principaux à la date de l'enquête, qui sont comme de

l'initiation, en 2004, a conduit à un échantillon de 16 000 logements, divisés en neuf parties égales. L'une de ces parties a été interrogée une seule fois (en 2004), une autre deux fois (2004 et 2005), une autre trois fois (2004, 2005 et 2006), etc. En régime de croisière, un panel donné sera interrogé neuf années de suite. Durant la phase d'initiation, qui s'achèvera en réalité en 2012 avec la sortie du neuvième et dernier sous-échantillon issu du tirage de 2004, les sous-échantillons seront évidemment sollicités

Les schémas suivants synthétisent les deux approches. Le rectangle du haut symbolise la population complète à t et celui du bas la population complète à $t + 1$. La partie « moins » représente les morts au sens large (décès, émigration, passage de l'individu en communauté,...) et la partie « plus » représente les naissances au sens large (nouveau-nés, immigration, entrée dans le champ par le franchissement d'un seuil d'âge,...). La partie grisée représente, à chaque date, la population d'inférence.

(l'absence de non-réponse). On peut s'intéresser en particulier à deux types de paramètres : des taux annuels X'_i (ou leurs satellites), ou des évolutions de totaux $\Delta_{i,t+1}$ entre deux années données, consécutives ou non. Pour simplifier, on considérera désormais qu'il s'agit de différences de totaux entre deux années consécutives. Quand on parle d'évolutions, il faut préciser les populations d'intérêt qui entrent en jeu. Il y a alors deux façons de voir les choses : soit on raisonne sur des populations évoluant avec le temps et l'approche est dite transversale, soit on raisonne à population fixe et l'approche est dite longitudinale. Si on note Ω_i la population complète du champ de l'enquête l'année i , le total annuel à l'année i est donné par $X'_i = \sum_{j \in \Omega_i} X'_{ij}$, où X'_{ij} est une variable d'intérêt mesurée pour l'individu j . Lorsqu'on parle d'évolution, l'objectif peut être d'estimer la différence $\Delta_{i,t+1}$ entre le total X'_i à $i+1$ sur Ω_{i+1} et le total X'_i à i sur Ω_i , c'est-à-dire $\Delta_{i,t+1} = X'_{i+1} - X'_i$. Ceci correspond à une vision transversale. Sinon, l'objectif peut être d'estimer la différence $\Delta_{i,t+1}$ entre les totaux définis sur les unités communes aux populations $\Omega_{i,t+1}$ et Ω_i , les différences d'effectifs entre les deux populations s'expliquant par les unités entrant (naissances) et sortant (morts) de ces populations. Ceci correspond à une vision longitudinale. Soit $\Omega_{i,t+1} = \Omega_i \cup \Omega_{i,t+1}$ la population commune entre i et $i+1$. On définit alors $\Delta_{i,t+1}$ selon $\Delta_{i,t+1} = \sum_{j \in \Omega_{i,t+1}} (X'_{i+1,j} - X'_{i,j})$.

Chaque année, on dispose d'un échantillon d'individus tous panélistes dont les huit neuvièmes ont déjà été interrogés au moins une fois les années passées (en

2. Principes généraux

Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France

Pascal Ardilly et Pierre Lavallée¹

Résumé

L'Enquête européenne sur le revenu et les conditions de vie (*Statistics on Income and Living Conditions*, SILC) a remplacé le Panel européen à partir de 2004. Elle permet de produire des statistiques annuelles sur la répartition des revenus, ainsi que sur la pauvreté et l'exclusion sociale. Cette enquête longitudinale, dont la collecte a eu lieu pour la première fois en France en mai 2004, touche tous les individus de plus de 15 ans occupant les 16 000 logements tirés dans l'échantillon-maître et la base de sondage des logements neufs. L'enquête doit aussi fournir des estimations transversales de qualité.

Afin de limiter la charge des enquêtes, le plan de sondage préconisé pour SILC par Eurostat est un schéma rotatif basé sur quatre panels d'une durée de quatre ans chacun avec remplacement d'un panel tous les ans. La France a néanmoins choisi de porter la durée de ses panels à neuf années. Le plan de sondage rotatif permet de répondre aux besoins longitudinaux et transversaux de l'enquête. Cependant, il pose des défis en matière de pondération.

Après un rappel du contexte de l'inférence lorsqu'on pratique une enquête longitudinale, l'article traite des pondérations longitudinales et transversales, qui sont conçues de manière à produire des estimateurs approximativement sans biais.

Mots clés : Enquête longitudinale; panel; méthode du partage des poids; pondération longitudinale; pondération transversale.

1. Introduction

L'enquête SILC (*Statistics on Income and Living Conditions*) est une enquête européenne portant sur la mesure du revenu et sur l'évaluation des conditions de vie des personnes vivant en ménage ordinaire (on exclut donc les personnes vivant en communauté). Elle a remplacé, à partir de l'année 2004, le panel communautaire. Bien que l'enquête soit européenne, et donc sous la tutelle d'Eurostat, elle est menée de façon indépendante à l'intérieur de chaque état membre de l'Union européenne. Les états membres - comme ici la France - peuvent ainsi adapter le plan de sondage proposé par Eurostat afin de répondre à leurs besoins nationaux. Le traitement des données est aussi effectué à l'intérieur de chaque état membre, comme c'est le cas habituellement pour les enquêtes d'Eurostat au sein de l'Union européenne. Le présent article se restreint au cas de l'enquête SILC en France, mais il pourrait aussi concerner les autres états membres de l'Union européenne.

L'enquête SILC est une *enquête longitudinale*, menée en mai de chaque année, qui s'intéresse aux individus physiques beaucoup plus qu'aux ménages, et qui est effectuée en face-à-face auprès de toutes les personnes résidant dans les logements échantillonnés. Cette enquête peut se voir comme la version européenne de l'Enquête sur la dynamique du travail et du revenu (EDTR) menée par Statistique Canada (voir Lavallée 1995, ainsi que Lèvesque et Franklin 2000).

L'échantillon de l'enquête SILC est rotatif : chaque année, à partir de 2004, il est constitué par la réunion de neuf sous-échantillons panels, tirés dans des conditions identiques en régime stationnaire, pour partie dans l'échantillon-maître, pour partie dans la base de sondage des logements neufs (BSLN). L'échantillon-maître et la BSLN sont deux bases de sondage de logements construites respectivement à partir du Recensement de la population (RP) de la France et des données du Système d'information et de traitement automatisé des données élémentaires sur les logements et les locaux (SITADEL) (voir Ardilly 2006). Chaque panel entrant dans SILC réunit tous les individus résidant dans l'ensemble des logements tirés. L'interrogation de l'ensemble sélectionnés permet de produire des estimations à la fois au niveau des individus et des ménages de la population. Elle permet, de plus, d'optimiser les coûts de collecte en maximisant le nombre d'individus atteints pour chaque contact. Certaines estimations concernant néanmoins un champ réduit, défini par les personnes ayant 16 ans ou plus au 31 décembre de l'année d'enquête.

Chaque année, un sous-échantillon sort et un sous-échantillon entre pour le remplacer. La première année où l'enquête a eu lieu, soit 2004, chaque sous-échantillon comprenait 1 780 logements (à quelques unités près, du fait des procédures d'arrondi). À partir de la seconde année, donc des 2005, la taille du sous-échantillon entrant de l'année a été fixée à 3 000 logements. Notons que

¹ Pascal Ardilly, Division « Échantillonnage et traitement statistique des données », INSEE, Paris, France. Courriel : pascal.ardilly@stat.can.ca, Lavalée, Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), K1A 0T6 Canada. Courriel : pierre.lavallee@statcan.ca.

- Laaksonen, S., et Chambers, R. (2006). Survey estimation under informative non-response with follow-up. *Journal of Official Statistics*, 81-95.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue internationale de Statistique*, 54, 139-157.
- Lundström, S., et Särndal, C.-E. (2001). Estimation in the presence of nonresponse and frame imperfections. Statistics Sweden.
- Rizzo, L., Kalton, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.
- Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
- Virtanen, V., Pouta, E., Sievähnen, T. et Laaksonen, S. (2001). Luonnon virkistyskäytön kysymätuutkimuksen aineistot ja menetelmät. (Données et méthodes d'enquête sur l'utilisation récréative de la nature). Dans Luonnon Virkistyskäyttö (l'utilisation récréative de la nature). Finnish Forest Research Institute, 802.
- Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Le problème dont traite le présent document est fréquent dans les enquêtes. Un grand nombre d'enquêtes se déroulent par phases et des incohérences s'y présentent à cause des données manquantes et d'autres divergences. Un exemple bien connu sur le plan international est l'Enquête sociale européenne (ESE) qui comporte deux questionnaires, à savoir un questionnaire de base et un questionnaire supplémentaire. Il va sans dire que le nombre de répondants est moindre pour le second de ces questionnaires que pour le premier, d'où une certaine sélectivité, par exemple, si le second questionnaire est en corrélation positive avec l'action politique. La situation sera étrange aux yeux de l'utilisateur lorsqu'une estimation varie selon qu'elle vient d'un grand ou d'un petit ensemble de données, bien que portant sur la même variable et la même période.

Comme pour l'ESE, notre étude vise des données d'enquête recueillies en deux phases. Le taux de réponse à la seconde enquête a été bien moindre que dans l'ESE. L'effet de sélectivité est également plus marqué. En appliquant les modèles de la proportion à répondre, nous avons pu prévoir cette sélectivité, exploiter les résultats dans des corrections de pondération et, en dernier lieu, calibrer la somme des poids sur certains sommations agrégats connus de la population.

Cette stratégie vise à tirer tout le parti voulu de l'information auxiliaire disponible tant dans les registres qu'à d'autres sources exécutrices, de même qu'à la première phase de l'enquête au niveau des micropodonnées. Dans notre exemple, la seconde phase de l'enquête est double, mais il n'existe qu'une collecte de données. Il y a d'abord le fait de se porter volontaire pour la seconde phase et ensuite la participation effective à l'enquête. Nous avons examiné les deux phases séparément pour trouver des données intéressantes sur les mécanismes respectifs de réponse. De plus, les résultats de cette analyse ont servi à la repondération. À des fins de comparaison, nous avons regardé les deux phases à la fois, élaboré un modèle applicable et continué la repondération d'une manière analogue. Enfin, nous avons comparé les estimations. Il était quelque peu étonnant que les deux ensembles de résultats diffèrent si peu dans nos exemples. C'est une bonne nouvelle cependant, puisqu'il est plus facile de travailler avec une pondération « une étape », laquelle pourrait alors être introduite.

Nous proposons un certain cadre méthodologique pour la pondération « deux étapes », mais sans pouvoir affirmer quelle spécification serait la meilleure dans chaque cas. Notre méthode est d'une application plutôt facile, mais les avantages qu'elle offre dépendent dans une large mesure, bien sûr, de la disponibilité de bonnes données auxiliaires

Remerciements

L'auteur aimerait remercier le rédacteur en chef de Techniques d'enquête et les examinateurs anonymes de leurs observations précises et utiles.

Bibliographie

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duncan, K.B., et Stasny, E.A. (2001). Utilisation de scores de propension pour contrôler le biais de couverture dans les enquêtes téléphoniques. *Techniques d'enquête*, 27, 2, 131-141.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Ekholm, A., et Laaksonen, S. (1991). Weighing via response modelling in the Finnish household budget survey. *Journal of Official Statistics*, 7, 2, 325-337.

Fuller, W.A., Loughin, M.M. et Baker, H. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Laaksonen, S. (1999). Weighing and auxiliary variables in sample surveys. Dans "Enquêtes et Sondages. Méthodes, modèles, applications, nouvelles approches" (Eds., G. Brosier et A.-M. Dussaux). Dunod, Paris, 168-180.

Laaksonen, S. (2006a). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications, An International Journal*. Editeur: IOS Press, 1, 95-100.

Laaksonen, S. (2006b). Need for high quality auxiliary data service for improving the quality of editing and imputation. Dans *United Nations Statistical Commission, "Statistical Data Editing"*, 3, 334-344.

Aux yeux de bien des utilisateurs, le biais demeure excessif pour les variables *nautisme* et *cyclisme*. Nous pouvons le réduire, bien entendu, en ajoutant des variables auxiliaires au modèle. Jusqu'où pourrions-nous aller dans cette direction? Nous n'examinons pas la question plus avant. Par ailleurs, nos outils sont encore pires s'il s'agit de réduire le biais de telles variables en fonction de la seule seconde phase. Nous avons vérifié plusieurs de ces estimations et nous observons certains changements dans les estimations correspondantes, et ce, au même niveau que pour les variables *nautisme* et *cyclisme* à la figure 7. Dans ce cas, il est cependant impossible de vérifier le biais. Nous pouvons uniquement supposer, en nous fondant sur les exercices précédents, que les résultats sont moins biaisés que ceux qui correspondent à des valeurs encore moins bien corrigées.

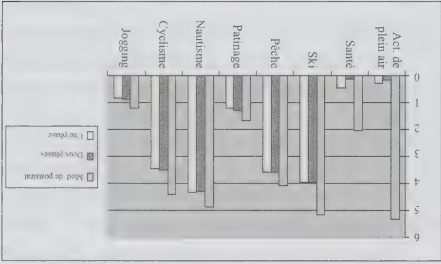


Figure 6 Biases des estimations en points de pourcentage de la réponse après les deux phases de poststratification et de la pondération de poststratification et des deux méthodes de la propension à répondre (pondération « deux étapes » et « une étape ») où les variables auxiliaires sont *activités de plein air et santé* (la méthode des échelons 2b et 3b au tableau 3 et la méthode « une étape » est directement fonction de la réponse en seconde phase (modèle 4b au tableau 3))

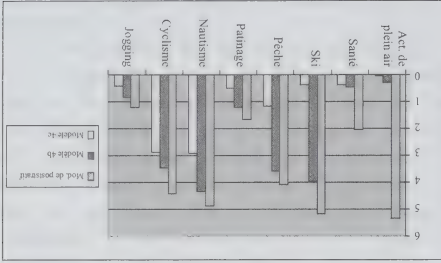


Figure 7 Biases des estimations en points de pourcentage de la réponse après les deux phases de poststratification et des deux méthodes de la propension à répondre où les variables auxiliaires sont *ski et pêche* (modèle 4c du tableau 3 en comparaison avec le modèle 4b)

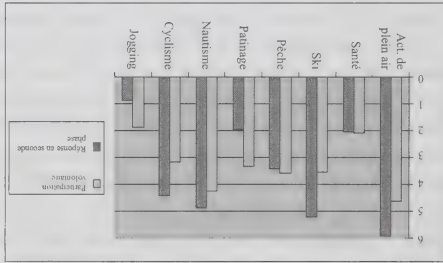


Figure 4 Biases des estimations en points de pourcentage en fonction de la pondération d'échantillonnage sans correction pour la participation volontaire et la réponse en seconde phase

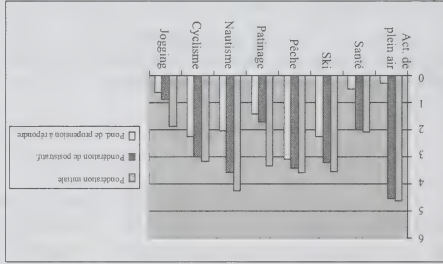


Figure 5 Biases des estimations en points de pourcentage de la participation volontaire en fonction de la pondération d'échantillonnage sans correction (symbole = « initial », du calage de poststratification et de la modélisation de propension à répondre où on emploie les variables auxiliaires *activités de plein air et santé* (modèle 2b au tableau 3))

servent de variables auxiliaires dans les modèles qui la supportent.

Les figures 6 et 7 visent les estimations de la dernière phase et sont donc les plus importantes. La figure 6 dégage la même conclusion que la figure 5, en ce sens que la technique de modélisation de propension à répondre se révèle supérieure à la technique de calage par poststratification, bien que toutes les différences ne soient pas hautement significatives statistiquement (en particulier pour la variable *jogging*). La différence entre la méthode « une étape » et la méthode « deux étapes » est plutôt ténue et le biais change d'une variable à l'autre. D'après cette étude, il est donc impossible de dire laquelle de ces deux spécifications est meilleure.

La figure 7 fait un certain nombre de comparaisons lorsqu'on ajoute deux variables au modèle de la propension à répondre. Les résultats sont tout à fait prévisibles, puisque le biais de ces estimations se trouve réduit, tout comme le biais de toutes les autres estimations dans une certaine mesure.

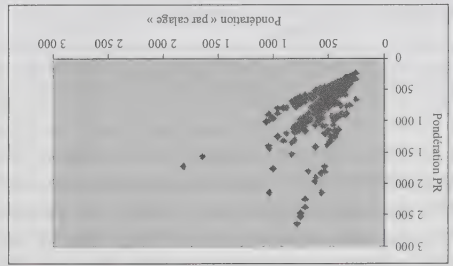


Figure 1 Diagramme de dispersion entre les deux pondérations de la participation volontaire

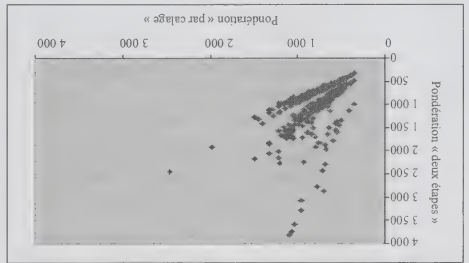


Figure 2 Diagramme de dispersion de la réponse en seconde phase entre la pondération par calage et la pondération « deux étapes » de la propension à répondre

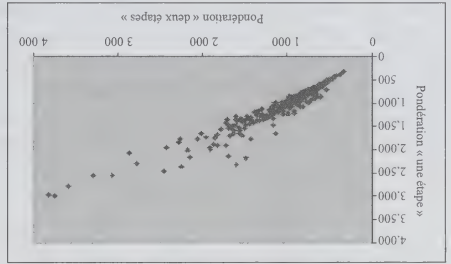


Figure 3 Diagramme de dispersion de la réponse en seconde phase entre les deux pondérations en modélisation de propension à répondre

4.3 Comparaison des estimations des paramètres

Nous n'avons pu faire d'études complètes de simulation avec des hypothèses variées et ainsi analyser quel type de méthode conviendrait le mieux dans chaque cas. Nous pouvons heureusement faire un grand pas dans cette direction en comparant les effets sur les estimations dans trois perspectives. Premièrement, nous modélisons la réponse et la participation volontaire en utilisant à la fois les variables X et un certain nombre de variables X_1 . Nous connaissons donc les meilleures estimations des paramètres par ces valeurs X_1 de la première enquête. Deuxièmement, nous ajoutons des variables auxiliaires X_2 mais en excluant

certaines. Nous connaissons néanmoins les « meilleures » valeurs dans ce cas et pouvons donc faire des comparaisons précises. Troisièrement, nous pouvons comparer des estimations sans information connue et, dans ce dernier cas, il est seulement possible de déduire les valeurs qui pourraient être les meilleures.

Nous présentons les données explicites selon les variables décrites à la section 2. À noter que nous n'avons pas jugé important de présenter les erreurs-types des diverses estimations, parce que nous nous concentrons sur les biais de ces dernières. On notera cependant que les erreurs-types se situent dans l'intervalle $[0,2-0,4]$ point pour l'ensemble de données première phase et dans l'intervalle $[0,3-0,5]$ point pour l'ensemble de données seconde phase (les valeurs sont les plus basses pour *santé*, deuxièmes pour *jogging* et les plus hautes pour *activités de plein air et ski*).

La figure 4 livre les résultats avec une pondération sans correction aucune (tel est souvent le cas dans la pratique malheureusement). Nous pouvons voir que les biais est appréciable dans la plupart des estimations; il est le plus faible pour *fogging*, qui n'est pas une catégorie de forte activité si on la compare à la variable *activités de plein air*. Par exemple, En règle générale, la plupart des utilisateurs n'aiment pas voir de tels biais importants et hautement significatifs statistiquement saut pour *fogging* en second

L'intention de l'utilisateur des données. À des fins de comparaison, nous présentons à la figure 5 les mêmes résultats non corrigés pour la participation volontaire qu'à la figure 4, mais nous ajoutons les estimations correspondantes par calage de poststratification et pour la modélisation de propension à répondre. Ce graphique indique clairement que la poststratification est source d'un certain avantage par rapport à toute solution hors correction. La méthode de la propension à répondre demeure la meilleure dans tous les cas cependant. Elle est extrêmement bonne pour les variables *santé et activités de plein air* qui

été à introduire. Avec un choix plus simple, celui d'une pondération non biaisée, on a une « étape » et il se trouve que les deux pondérations avec une telle pondération respectives à répondre de manière à mieux comprendre les causes des deux types d'état de données manquantes.

(les numéros de modèle aux tableaux 3 et 2 se correspondent et la variable de réponse et les ensembles de données sont les mêmes)

Variables explicatives et autres statistiques	Modèle 2b	Participation volontaire	Modèle 3b	Réponse des volontaires	Modèle 4b	Réponse en seconde phase	Modèle 4c	Réponse en seconde phase
---	-----------	--------------------------	-----------	-------------------------	-----------	--------------------------	-----------	--------------------------

Hommes	0,94	0,77	0,82	0,75
--------	------	------	------	------

	24 ans ou moins	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 ans et plus
4,92	0,52	1,30	0,51			
(4 07; 5 97)	(0 41; 0 65)	(1 12; 1 52)	(0 41; 0 64)			

25-34	3.83 (0.00, 6.06)	0.65 (0.00, 1.30)	1.46 (0.00, 2.92)	(3.18, 4.60) (0.52, 0.81)	(1.25, 1.70) (0.52, 0.81)	(0.52, 0.81) (0.00, 0.65)
-------	----------------------	----------------------	----------------------	------------------------------	------------------------------	------------------------------

45-54	2.59	0.85	1.56	1.55	1.15
	(2.20; 3.05)	(0.68; 1.06)	(1.34; 1.81)		(0.69; 1.06)
55-64	1.73	1.18	1.55	1.55	1.15

Variable	Mean	SD	95% CI	95% CI	95% CI
Age	58.1	10.2	57.9, 58.3	57.7, 58.5	57.5, 58.7
Gender	Male	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Marital status	Married	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Education	High school	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Income	\$10,000-\$20,000	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Health status	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Smoking status	Non-smoker	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Alcohol consumption	Non-drinker	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Exercise frequency	Regular	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Stress level	Low	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Sleep quality	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Dietary habits	Healthy	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Family size	2-3 children	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Work status	Employed	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Comorbidities	None	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Medication use	None	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare access	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare utilization	High	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare costs	Low	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare satisfaction	High	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare quality	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare access	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare utilization	High	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare costs	Low	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare satisfaction	High	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0
Healthcare quality	Good	1.0	1.0, 1.0	1.0, 1.0	1.0, 1.0

quest central	quest periferica	quest periferica	quest central
1.71.254)	(0.93; 1.43)	(1.29; 1.78)	(0.93; 1.43)
2.09	(1.05; 1.15)	1.52	(1.05; 1.15)
(1.04; 2.20)	(1.17; 1.50)	(1.17; 1.50)	(1.04; 2.20)
1.16	(0.93; 1.43)	(0.93; 1.43)	(0.93; 1.43)

Activités de plein air	3,43 (2,71)	1,43 (1,07)	1,93	1,77
	3,04	1,24	1,93	1,77
	(0,98; 1,41)	(0,91; 1,42)	(1,00; 1,58)	(0,92; 1,43)

1.36
(3.24; 1.92)

Nombre d'observations	10 666	8 481	10 666	(1,38; 1,17)
-----------------------	--------	-------	--------	--------------

Tableau 4 Statistiques descriptives de différentes valeurs de pondération d'échantillonnage (PR = propension à répondre)

Pondération	Phase	Nombre d'unités	Moyenne	Asymétrie	$1 + cv^2$
-------------	-------	-----------------	---------	-----------	------------

Pondération	Phase	Nombre d'unités	Moyenne	Asymétrie	1 + cv ²
Poids de sondage	Zéro	12 658	308	0,94	1,30
Poids de base par calage	Premier	10 666	365	1,30	1,39
Poids en calage	Volontaires	8 481	460	2,52	1,63
Pondération de propension à répondre (PR), modèle 2b	Volontaires	8 481	460	4,60	1,82
Poids par calage	Deuxième	5 480	712	1,64	1,62
Pondération PR « deux étapes », modèles 2b et 3b	Deuxième	5 480	712	3,60	1,84
Pondération PR « une étape », modèle 4b	Deuxième	5 480	712	2,56	1,80

sondage ne peuvent intervenir dans nos comparaisons, car nous ne disposons pas de données sur les variables X pour l'échantillon initial. On peut cependant voir que c'est pour cette pondération que la variation relative mesurée ici est la plus faible ($1 + cv^2$, cv étant le coefficient de variation). Cette formule sert également d'approximation de l'effet de plan. Rizzo et coll. (1996) se servent aussi de cet indicateur lorsqu'ils comparent les pondérations.

Les changements ne sont pas considérables pour la première phase, c'est-à-dire lorsqu'on passe du poids de sondage au poids de base (sauf pour la moyenne liée à une taille d'échantillon qui décroît), mais aux deux phases qui suivent, l'effet de plan est plus marqué. On constate en outre que la variation est moindre pour les deux pondérations par calage que pour la pondération basée sur la modélisation de propension à répondre. Pour chaque pondération, la distribution est asymétrique à droite; elle l'est le moins pour les poids de sondage, bien sûr. On s'étonnera quelque peu que l'asymétrie soit la plus grande pour la pondération de la participation volontaire. On trouvera aux figures 1 à 3 plus de détails sur les distributions en question et les différences entre les pondérations.

La figure 1 illustre bien que certaines valeurs de pondération s'accroissent nettement à cause de la modélisation de propension à répondre (modèle 2b). Il est possible de voir en détail quels types d'unités sont sous les courbes à valeurs de pondération largement en croissance. Ainsi, derrière les courbes du côté gauche où les valeurs sont de plus de 700 pour la pondération par la propension à répondre, on trouve des gens qui ne sont pas en santé et qui ne raffolent pas des activités de plein air, mais qui n'en figurent pas moins au fichier de la participation volontaire. De même, on découvre d'autres groupes intéressants en se reportant aux résultats des estimations par modélisation. La majeure partie des courbes sont cependant concentrées et, par conséquent, on peut s'attendre à moins de changements dans les estimations que la no la pondération varierait plus amplement.

La dispersion est quelque peu plus grande à la figure 2 qu'à la figure 1, mais le profil demeure le même. On peut donc trouver des sous-groupes intéressants derrière des courbes distinctes.

À la figure 3 enfin, nous comparons les deux pondérations de seconde phase. Ce diagramme de dispersion est considérablement des deux précédents, car la relation est plutôt linéaire. Les valeurs maximales sont plus élevées pour la pondération « deux étapes » que pour la pondération « une étape », mais les valeurs de pondération « une étape » sont dans bien des cas nettement supérieures. Ainsi, les gens qui ne sont pas en santé et qui se livrent cependant à des activités de plein air reçoivent une valeur « une étape » relativement élevée, mais il n'y a pas d'effet net de l'âge. Par ailleurs, les membres des tranches d'âge supérieures qui

Nombre de résultats intéressants sont à relever dans ces modèles de comportement échelonné de données manquant. Les résultats de la première enquête sont plutôt normaux; par exemple, les hommes répondent plus médiocrement que les femmes aux deux phases. La propension à répondre est également plus faible dans le sud que dans le reste du pays. Les différences entre tranches d'âge sont quelque peu étonnantes, puisque les groupes d'âge intermédiaires sont ceux qui répondent le moins bien.

Les estimations de la participation volontaire sont différentes. Dans l'est central et le nord, les gens sont le moins disposés à participer à la seconde phase, mais les résultats de réponse une fois que les gens se sont portés volontaires ne diffèrent pas nettement. Selon l'âge, il semblerait que les gens plus jeunes sont plus disposés à participer, mais sans très bien répondre par la suite. Ainsi, les gens plus âgés seraient en ce sens plus prêts à s'engager à participer, mais on peut clairement voir que les plus âgés seront sous-représentés hors de toute correction.

Si nous considérons les deux premières variables auxiliaires internes (tableau 3), nous pouvons observer que les gens qui ne sont pas relativement en santé (variable *santé*) et qui ne s'adonnent pas activement aux loisirs en milieu naturel (*activités de plein air*), ne sont pas disposés non plus à recevoir tout nouveau questionnaire, à en juger par le rapport de probabilité très élevé dans ce cas. Fait intéressant, les rapports de probabilité sont proches pour la variable *santé* et la participation volontaire. En d'autres termes, la personne qui n'est pas en santé ne sera pas très encline à se porter volontaire, mais si elle le fait, sa réponse ressemblera à celle de la personne en santé. Une tendance semblable se dégage en ce qui concerne la variable *activités de plein air*. Il convient de noter que les dominaines « non en santé » et « non en plein air » ne sont pas très étendus et que, bien que jouant un grand rôle dans la modélisation de la propension à répondre, leur incidence sur les estimations obtenues à la fin n'est pas considérable (section 4.3).

Si on ajoute les deux autres variables auxiliaires internes, celles du *ski* et de la *pêche*, la même sélectivité est constatée, mais elle n'est pas aussi appréciable. En conclusion, on discerne clairement que le mécanisme de réponse à la seconde enquête serait très peu « non informatif ». Il faut s'attendre à certains effets, par conséquent, sur la pondération et les estimations d'enquête, ce que nous examinerons dans les deux sous-sections qui suivent.

Comme nous l'avons exposé, nous fournissons plusieurs poids. Le tableau 4 les récapitule avec les statistiques descriptives de manière à expliquer les changements considérés à chaque correction de pondération. Les poids de

4.2 Comparaison entre les poids

4. Résultats empiriques

Cette section décrit les résultats de l'application de différentes méthodes. Nous exposons les résultats de divers modèles de réponse, puis comparons les valeurs de pondération entre elles pour enfin confronter un certain nombre d'estimations des paramètres issues des différentes techniques.

4.1 Modélisation de la participation volontaire et de la réponse

Pour comprendre à fond le comportement de données manquantes (par absence de réponse et de participation volontaire) aux trois stades dégagés pour l'enquête, nous présentons au tableau 2 des résultats fondés sur les variables auxiliaires X_i disponibles à chaque phase (dans la pratique, nous comptons de ses effets dans cette analyse, puisqu'il ne s'agit pas là d'un sujet d'intérêt premier pour notre propos).

Il n'y a pas que les techniques clés dont nous venons de parler, à la prochaine section, nous livrons aussi en pondération w_i nos *milleures estimations possible* pour les paramètres qui sont connus (variables X_i).
De plus, nous mettons nos résultats bien précis en comparaison par rapport au seul calage de poststratification sans modélisation (symbole « *cal* » dans les sections qui restent de notre exposé). Il s'agit là d'une manière tout à fait normale d'aborder le problème de la pondération (c'était le « style maison » avant cette proposition méthodologique). Il convient toutefois de noter que, si un modèle de réponse comprend seulement les variables (et les mêmes catégories) de poststratification, la pondération en modélisation de la propension à répondre correspond parfaitement à celle que l'on obtient par la technique de poststratification.

Tableau 2
Régressions logistiques par les trois variables explicatives communes des trois phases, c'est-à-dire de la réponse en première phase et de la participation volontaire et de la réponse réelle en seconde phase. Nous présentons les estimations sous forme de rapports de probabilité; les intervalles de confiance à 95 % figurent entre parenthèses

Variables explicatives	Modèle 1	Modèle 2a	Modèle 3a	Modèle 4a
Réponse en première phase				
Participation volontaire				
Réponse des volontaires				
Réponse en seconde phase				
Sexe (réf. : sexe féminin)	0,71	0,84	0,75	0,77
Tranche d'âge (réf. : 65 ans et plus)	(0,65; 0,78)	(0,76; 0,93)	(0,68; 0,83)	(0,71; 0,83)
24 ans ou moins	1,00	5,57	0,51	1,49
25-34	(0,83; 1,21)	(4,65; 6,68)	(0,41; 0,64)	(1,28; 1,73)
35-44	(0,79; 1,15)	(4,00; 4,81)	(0,52; 0,81)	(1,49; 2,00)
45-54	0,85	4,08	0,64	1,62
55-64	(0,71; 1,02)	(3,46; 4,81)	(0,52; 0,80)	(1,40; 1,88)
45-54	0,89	3,16	0,86	1,82
55-64	(0,74; 1,07)	(2,71; 3,69)	(0,69; 1,06)	(1,58; 2,10)
55-64	1,18	2,05	1,15	1,75
Région (réf. : nord)	(0,96; 1,45)	(1,74; 2,41)	(0,90; 1,47)	(1,49; 2,04)
sud-est	0,55	2,12	0,96	1,35
sud-ouest	(0,46; 0,66)	(1,79; 2,50)	(0,79; 1,16)	(1,17; 1,55)
sud-ouest	0,76	1,83	1,04	1,35
sud-ouest	(0,64; 0,91)	(1,57; 2,14)	(0,86; 1,25)	(1,18; 1,55)
ouest central	1,14	2,14	1,16	1,56
ouest central	(0,91; 1,42)	(1,77; 2,59)	(0,93; 1,43)	(1,33; 1,83)
est central	0,96	1,20	1,15	1,19
est central	(0,78; 1,18)	(1,01; 1,44)	(0,92; 1,43)	(1,02; 1,40)
Nombre d'observations	12 554	10 666	8 481	10 666
-2 log L	10 904	10 296	8 569	14 618

recoupement des trois variables X de la *branche d'âge*, du *sexe* et de la *région*, le nombre total de cellules est de $6*2*5 = 60$). La technique est simple :

$$g_h = \frac{\sum_h w_k}{\sum_h w_k / p_k}$$

Comme nous l'avons mentionné, on obtient en Finlande une haute qualité pour ce genre d'aggrégats de poststratification, mais non pas nécessairement pour les autres. Comme la seconde phase est double, nous avons les trois possibilités suivantes de modélisation qui sont toutes exploitées à la section 4 :

- a) modèle relatif à la participation volontaire;
- b) modèle relatif à la réponse des personnes qui se sont portées volontaires (modélisation « deux étapes »);
- c) modèle de réponse en une étape (modélisation « directe » ou « une étape »).

À noter que, combinées, les phases a) et b) donnent les valeurs de pondération du fichier B, d'où la formulation suivante (vol = participation volontaire; p_1 = probabilité estimée de réponse à la phase 1; p_2 = probabilité estimée de réponse à la phase 2; facteurs d'échelle g_1 et g_2 respectivement) :

$$\begin{aligned} \text{Phase 1 : } w_k(\text{vol}) &= \frac{p_{1k}}{w_k} g_{1k}, \\ \text{Phase 2 : } w_k(\text{res}) &= \frac{p_{2k}}{w_k(\text{vol})} g_{2k}. \end{aligned}$$

Il a fallu utiliser les bons poids de sondage pour chaque tâche de modélisation. Pour les modèles a) et c), ce sont les nos tests de comparaison, nous avons également modélisé la réponse en première phase et utilisé les poids de sondage. En pondérant dans la modélisation, nous obtenons des estimations plus fidèles, puisques nous lachons de rendre notre analyse représentative de la population cible. Parfois, la pondération est d'une grande incidence comme dans les enquêtes auprès des entreprises où, souvent, les valeurs de pondération sont plus variables que dans les enquêtes types auprès des ménages. Dans le cas qui nous occupe, les résultats ne diffèrent pas outre mesure selon que les modèles étaient en pondération ou non, bien qu'une pondération s'impose en pareil cas (c'est ce que démontrent nettement Laaksonen et Chambers (2006) où l'incidence de la pondération est appréciable; Rizzo et coll. (1996) à la section 4 les données empiriques (estimations avec leurs erreurs-types et les probabilités de réponse) pour les solutions en pondération.

À la section 4, nous faisons deux types de comparaisons (i) en fonction de l'information connue de la première phase et (ii) sans nous reporter à l'information connue. Dans l'un et l'autre cas, nous pouvons heureusement fort bien savoir si nous avons réussi à réduire le biais, puisque nous disposons des estimations « vraies » (ou des meilleures estimations possibles). Ajoutons que nous analysons un certain nombre de variables qui figurent seulement dans le second questionnaire, mais qu'il est impossible d'établir le degré d'efficacité de chaque méthode dans ce cas. Nous ne présentons pas non plus les derniers résultats en détail, mais on sait par observation que le tout se comporte comme dans le deuxième cas.

3. Modélisation de la propension à répondre : méthode et calage

Notre étude comprend trois phases avec les spécifications suivantes de pondération :

D'abord, on met au point une bonne pondération calibrée pour les répondants de première phase avec les variables *région, sexe, tranche d'âge et saison* (voir aussi la section 1). Le symbole w_k désignera le poids de sondage du répondant k . Ces poids sont basés sur le calage et sont aussi appelés poids « de base ». À noter que, préalablement, il y a eu des poids de sondage de crées pour l'ensemble de données en fonction d'un plan d'échantillonnage aléatoire stratifié. Nous disposons de telles valeurs des poids de sondage pour les non-répondants en première phase.

En second lieu, nous modélisons les probabilités de participation volontaire et de réponse à l'aide de la fonction de combinaison la plus courante (la fonction logit n'est pas nécessairement la meilleure de ces fonctions, nous le savons depuis (2006a), mais c'est elle que nous utilisons ici). Dans ce cas, $\text{logit} = \log(\pi/(1 - \pi))$, où π est la probabilité binaire de réponse (1 = participation volontaire et 0 = non-participation volontaire ou 1 = réponse et 0 = non-réponse). Les variables explicatives employées sont les variables X et un certain nombre de variables X_1 considérées comme « bonnes ». Ce modèle donne les probabilités prévues de réponse p_k qui servent de la manière suivante à la pondération dans chaque cas :

$$w_k(\text{res}) = \frac{p_k}{w_k} g_k$$

Dans ce cas, $g_c = a$ est un facteur d'échelle qui étalonne en fonction de certains aggrégats connus au niveau c . Il existe plusieurs solutions de rechange à ce mode de calage, mais ce que nous décrivons pourrait être considéré comme un calage type. Dans notre étude, nous utilisons les aggrégats h en poststratification (c'est la cellule de

Bien entendu, nous comparons les résultats obtenus par diverses techniques. À la section 2, nous décrivons sommairement nos questionnaires et nos ensembles de données; à la section 3, nous détaillons les principes qui sous-tendent nos méthodes; à la section 4, nous présentons des données de comparaison et, à la section 5, nous tirons nos conclusions.

2. Principes sous-tendant les ensembles de données

Les données viennent d'une enquête spéciale menée auprès de Finlandais de 15 à 74 ans (pour plus de renseignements à ce sujet, voir Virtanen, Pouta, Sievänen et Laaksonen 2001). Le thème abordé est celui des activités en période de loisirs et notamment des activités et passe-temps de plein air. Il y a d'abord eu une interview sur place assistée par ordinateur (IPAO) où on a posé des questions diverses sur les loisirs et les hobbies des gens (cyclisme, motocyclisme, randonnée pédestre, jogging, natisme, natation, chasse, pêche, photographie d'observation de la nature, ski, patinage, équitation, etc.). Dans tous les cas, la période de référence était l'année précédente. À la fin de cette enquête, on a demandé aux répondants s'ils seraient disposés à recevoir un questionnaire spécial d'enquête postale qui leur permettrait de décrire plus en détail certaines de ces activités. L'enquête en question aurait lieu quelques semaines après.

La période d'enquête a été de deux ans (1998-2000), le but étant d'alléger le fardeau de réponse et les charges d'interview. Autre facteur : les activités visées étaient saisonnières dans une certaine mesure, et on pouvait s'attendre à ce que les réponses subissent l'influence des saisons (on ne répond pas, par exemple, de la même manière à des questions sur le ski si on est en été ou en hiver). Après correction de surdénombrement (104 unités), l'échantillon initial s'est établi à 12 554 personnes.

Nous avons choisi les variables binaires suivantes pour l'analyse présentée à la section 4 : *activités de plein air* (l'enquête exerce régulièrement certaines de ces activités dans la nature), *santé* (l'enquête est-il assez en santé pour mener de telles activités?), *ski*, *pêche*, *patinage*, *nautisme*, *cyclisme* et *jogging*. Dans tous les cas, nous attribuons la valeur 1 si on a exercé l'activité en question l'année précédente et la valeur 0 si on ne l'a pas fait. Toutes ces variables étaient comprises dans le questionnaire de première phase; nous savions donc à quoi nous attendre après les deux phases. Précisons que notre ensemble de données comportait des variables plus complexes, mais que nous avons jugé bon de simplifier pour que les résultats soient d'une interprétation plus facile. Les grandes conclusions seraient les mêmes si notre choix avait été autre.

Établi en conséquence la pondération de base des répondants de la première phase en ajoutant la variable *saison* ($4 \times 2 = 8$ catégories sur deux ans) à nos variables auxiliaires. Dans les enquêtes réalisées en Finlande, on emploie rarement la variable « *saison* », mais le caractère « saisonnier » de notre enquête rendait cette utilisation nécessaire (voir la section 2). Nous n'examinerons pas cet aspect en détail. Les trois premières variables reviennent souvent dans nos enquêtes auprès des personnes, cette information se prêtant d'emblée à la validation par un registre de la population mis à jour. Cela étant, nous supposons disposer des meilleures estimations possible pour les répondants de la première phase lorsque nous utilisons cette pondération par calage. De toute manière, nous n'avons pas d'autre accès à des sources externes d'information susceptibles de nous être utiles.

Il est possible de se reporter aux estimations obtenues des répondants de première phase pour le calage de seconde phase. C'est une stratégie qui n'est pas difficile à appliquer en soi, mais on devrait inclure dans cette application toutes les variables X et autant de variables Y que possible. Il faut également trouver pour une telle stratégie des niveaux d'ensemble ou de domaine qui soient le plus précis possible, tâche ardue que nous n'entreprendons pas dans cette étude. Rien ne garantit non plus que les estimations d'autres aggregates seront suffisamment exemptes de biais (les résultats de Laaksonen (1999) livrent certaines indications à l'appui de cette conclusion).

La stratégie que je propose est plus simple et fonctionne sans problèmes techniques pour tous les domaines, bien qu'on n'ait nullement la garantie, bien sûr, qu'un éventuel biais sera largement réduit pour l'ensemble des domaines. Je n'ai donc opté pour aucune stratégie avancée de calage, bien que la chose soit praticable. J'espère que d'autres auteurs en montreront les avantages possibles. Une référence utile à ce propos est Dupont (1995), qui parle de calage de données d'enquête en deux phases, mais sans fournir d'information empirique à cette fin. À noter donc que je procède par calage, mais sans prendre quelque chose de très avancé (voir la section 3).

La méthode que je propose fait largement appel à une modélisation de propension à répondre qui a été employée avec succès dans d'autres cas (voir, par exemple, Ekholm et Laaksonen (1991), Laaksonen (1999), Duncan et Shany Kallton et Brick (1996) rappelle assez notre enquête en deux phases, mais il s'agit d'une enquête longitudinale. Le cadre méthodologique appliqué par ces auteurs à un certain nombre de caractéristiques semblables. Une grande différence est le mécanisme de réponse qui intervient ici en deux phases (participation volontaire et réponse à la deuxième phase). Nous analysons ces stades séparément nous aussi.

élevé (10 666 unités sur 12 554, soit un pourcentage approximatif de 85 %) à la première enquête. On a observé une certaine déperdition d'effectif parce que les répondants n'étaient pas tous disposés à participer à la seconde enquête (il ne reste plus alors que 68 % de l'échantillon initial). Comme le taux de non-réponse est plutôt élevé en seconde phase (malgré le caractère volontaire de la participation des répondants), le sous-échantillon restant n'est plus que de 44 % de l'échantillon initial. Nous disposons de trois ensembles de données qui suivent à des fins d'analyse :

- A. Répondants en première phase avec les variables d'enquête X_1 .
- B. Répondants en seconde phase avec les variables d'enquête X_1 .
- C. Répondants tant en première qu'en seconde phase avec les variables d'enquête X_1 et X_2 .

La plupart des utilisateurs recevront deux fichiers, A et B , qu'ils pourront réunir en un fichier C s'ils peuvent trouver un identificateur commun. À quoi s'attend l'utilisateur au reçu des deux fichiers? Il s'attend, bien sûr, à ce que les estimations d'un même paramètre soient identiques dans les deux fichiers, c'est-à-dire que les résultats soient homogènes, sachant que l'estimation d'un paramètre par un fichier C de moindre taille est moins précise qu'une estimation par un fichier plus complet. En principe, il est possible d'imputer les valeurs manquantes des variables X_2 mais nous pensons qu'il est impossible de bien le faire, aussi vaut-il mieux pour nous procéder par pondération, le but étant de trouver des corrections de pondération du fichier B pour la meilleure analyse possible sur les variables X_1 et X_2 .

Plusieurs stratégies sont applicables en pareil cas. Les aspects généraux d'intérêt sont présentés entre autres par Kalton et Kasprzyk (1986), Little (1986), Särndal, Swensson et Wretman (1992), Fuller, Loughin et Baker (1994), Wu et Sitter (2001) et Lundström et Särndal (2001). Si nous posons que l'état de données manquantes ne dépend

que du plan d'échantillonnage, nous pouvons pondérer les fichiers A et B en ce sens. En cas d'échantillonnage aléatoire stratifié par exemple, la même stratification vaudrait naturellement pour les deux phases. En cas de poststratification, une stratégie analogue pourrait s'employer. Dans l'exemple précis de notre enquête, la base de sondage comportait le premier groupe de renouvellement en 12 mois de l'EPA finlandaise. Chaque mois, l'échantillon est prélevé au hasard. Notre EPA se fait par échantillonnage aléatoire simple, mais à cause de la non-réponse, on répondrait par une technique type de calage (Deville, Särndal et Sautory 1993) en prenant le sexe, la tranche d'âge (six catégories) et la région (cinq catégories) comme variables auxiliaires. C'est ce que nous allons appeler les poids de sondage dans la suite de notre exposé. Nous avons

la poste ou par courrier électronique. Un exemple récent en est l'Enquête sociale européenne où le questionnaire supplémentaire porte plus particulièrement sur les valeurs de la vie (voir www.europeansocialsurvey.com). Bien sûr, ce ne sont pas tous les répondants de la première interview qui vont remplir le questionnaire qui suit.

Les questions posées en seconde phase n'abordent pas nécessairement le même thème que celles du questionnaire de base. Une autre stratégie courante est de prendre d'abord un questionnaire général sur un thème et de revenir en seconde phase avec des questions plus détaillées sur le même thème. Il peut aussi y avoir une certaine réaction de la première phase pour le second questionnaire et même un échantillon qui dépend de la distribution des variables cibles de la première enquête (c'est là un exemple d'échantillonnage adapté). On procède fréquemment de la sorte lorsqu'on n'a pas vraiment d'expérience avec ce type d'enquête. La première phase sert aussi d'enquête pilote. Avec ce qu'on appelle les échantillons-maîtres, on est proche de l'idée que l'enquête de première phase (ce qui peut comporter un échantillonnage à des sources admistratives comme microrécensement dans certains pays) puisse servir à élaborer une base de sondage appropriée. Dans ce cas, les variables de l'échantillon-maître sont plutôt limitées, ne livrant habituellement que des données de fait ou des renseignements généraux.

Avec les échantillons-maîtres, on vise à ce que la base de sondage obtenue représente bien la population visée. En cas d'échantillonnage, la base de sondage pourrait provenir les données auxiliaires nécessaires à la vérification, à l'imputation et à la pondération pour l'enquête bien réelle qui doit suivre (enquête de seconde phase). Chaque enquête réelle exploite donc un sous-échantillon de l'échantillon-maître. Nous examinerons ici un cas plus complexe qu'illustre le tableau 1.

Tableau 1 Illustration de l'échantillon initial avec trois ensembles de données en suivi

Echantillon avec variables auxiliaires	X (sexe, âge, région et saison)	Poids selon le plan pour 12 554 unités sans surdénom-brement
Répondants en première phase	X_1 (santé, activités de plein air, etc.)	Poids pour base de 10 666 unités
Volontaires en seconde phase	X_1 avec variables	Poids pour 8 481 unités
Répondants en seconde phase	X_2 (pâtinage, nautisme, etc.)	Poids pour 5 480 unités

On applique d'abord une procédure type d'échantillonnage avec un certain nombre de variables auxiliaires X . En l'occurrence, on a obtenu un taux de réponse relativement

Pondération de données d'enquête recueillies en deux phases

Seppo Laaksonen¹

Résumé

L'état de données manquantes peut prendre diverses formes. Dans cet exposé, nous nous attachons à la non-réponse des unités et tentons de corriger cet état par une pondération appropriée. Le cas empirique que nous présentons vise l'échantillonnage à deux phases. En première phase, on a fait enquête auprès d'un grand échantillon à l'aide d'un questionnaire plutôt général. Au terme de cette phase, l'intervieweur a demandé aux répondants s'ils désiraient participer à une seconde phase où, avec un questionnaire plus détaillé, on se concentrerait sur un certain nombre de thèmes venant de la première phase. Cette procédure crée trois mécanismes de données manquantes. La difficulté est de savoir comment pondérer le plus exactement possible les répondants de la seconde phase par souci de cohérence des données issues des deux phases. Nous analyserons d'abord les différences de données manquantes portées à un tel scénario en trois temps en nous reportant à des données d'enquête auprès des personnes et nous comparerons ensuite divers modes de pondération. Notre recommandation est d'utiliser toutes les données auxiliaires disponibles le mieux possible. On obtient un bon résultat en mettant les deux méthodes classiques avec d'abord une pondération de propension à répondre et ensuite un calage sur la distribution connue de la population.

Mots clés : Calage; variables auxiliaires internes et externes; méthode de modélisation de la propension à répondre; sous-échantillonnage sélectif.

1. Introduction

(1) Une enquête peut être à deux (ou à trois à certains égards) étapes ou phases. La première phase est celle de l'enquête type à laquelle répondent un certain nombre d'unités. En seconde phase, nous gardons seulement dans la base de sondage les répondants qui sont disposés à répondre à un questionnaire plus détaillé. Il faut d'abord distinguer les répondants en première phase qui disent être disposés à participer de ceux qui, dans ce même groupe, répondront effectivement au second questionnaire. Ce sous-groupe se trouvera donc à remplir les deux questionnaires.

(ii) Dans nos corrections après enquête, nous aurons la possibilité d'exploiter des variables auxiliaires tant externes qu'internes en seconde phase. Les variables internes viendront de la première enquête.

Notre propos est l'examen de la seule pondération, mais certaines des idées que nous avançons sont aussi applicables à l'imputation. Ce que nous décrivons n'a pas beaucoup été utilisé dans le cas des enquêtes transversales, mais le même problème s'est souvent posé. Il arrive fréquemment, par exemple, que l'intervieweur interroge directement l'enquêté en premier lieu et que, à la fin de cette interview initiale, il demande au répondant de remplir lui-même un questionnaire par la suite. Si l'intéressé est d'accord, l'intervieweur lui remet immédiatement le questionnaire à remplir ou l'envoie ultérieurement. Dans l'une ou l'autre de ces situations, il recevra les réponses par

L'enquête type se fait en une étape ou phase. Dans un tel cas, on aura sélectionné en premier lieu des unités en appliquant un certain plan d'échantillonnage et on aura tâche de prendre contact avec les personnes choisies et de les interviewer le mieux possible. Ce faisant, on aura créé une certaine non-réponse à un degré variable ou encore d'autres formes de données manquantes ou lacunaires. Pour corriger un état de données manquantes, on recourt habituellement à des méthodes postenquête de correction plus ou moins avancées qui font intervenir les variables auxiliaires disponibles à diverses sources (voir, par exemple, Laaksonen 1999, ou une version élargie dans Laaksonen 2006b, et Lundström et Särndal 2001). Il reste que les données servant à la pondération seront normalement tirées de registres ou d'autres sources administratives ou enquêtes. On pourra parler de variables auxiliaires *externes* à distinguer des variables *internes* de même nature; elles seront internes en ce sens que l'information viendra de la même enquête ou de celle qui l'a précédée.

Les variables auxiliaires internes servent en particulier à l'imputation si des valeurs manquent. Elles sont aussi beaucoup utilisées dans les enquêtes longitudinales s'il y a réponse à un cycle d'enquête et non-réponse à un autre. Dans ce cas, l'information interne peut servir à la fois à la repondération et à l'imputation.

Notre exposé ne porte pas sur une enquête type comme celle que nous venons d'évoquer. Il vise deux caractéristiques :

- Renssen, R.H., et Nieuwenbroek, N.J. (1997). Alliging estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.
- Renssen, R.H., Kroese, A.H. et Willeboordse, A.J. (2001). Alliging estimates by repeated weighting. Rapport, Central Bureau of Statistics, Pays-Bas.
- Rizzo, L., Kallott, G. et Brick, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.
- Rueda, M., Martinez, S., Martinez, H. et Arcos, A. (2007). Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, 137, 435-448.
- Samdal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Planning and Inference*, 7, 155-170.
- Samdal, C.-E., et Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.
- Samdal, C.E., Swensson, B. et Wetman, J. (1992). *Model-assisted Survey Sampling*. New York : Springer-Verlag.
- Samdal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York : John Wiley & Sons, Inc.
- Samdal, C.-E., et Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Statistics Sweden : Research and development - methodology report 2007:2*.
- Singh, A.C., et Mohl, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- Singh, S., Horn, S. et Yu, F. (1998). Estimation de la variance de l'estimateur général de régression : approche de calage à niveau élevé. *Techniques d'enquête*, 24, 43-52.
- Skinner, C. (1998). Calibration weighting and non-sampling errors. *Proceedings International Seminar on New Techniques for Statistics*. Sorrento, novembre 4-6, 1998, 55-62.
- Statistique Canada (1998). Methodology of the Canadian Labour Force Survey. Statistique Canada, Division des méthodes d'enquêtes auprès des ménages. Ottawa : Minister of Industry, catalogue no. 71-526-XPB.
- Statistique Canada (2003). Quality Guidelines (quatrième édition). Ottawa : Minister of Industry, numéro de catalogue 12-539-XIE.
- Steel, D.G., et Clark, R.G. (2007). Estimation par la régression au niveau de la personne et au niveau du ménage dans les enquêtes ménages. *Techniques d'enquête*, 33, 59-69.
- Snukel, D.M., Hildengton, M.A. et Samdal, C.-E. (1996). Estimation des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.
- Thøberge, A. (1999). Extension of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.
- Thøberge, A. (2000). Calage et poids restreints. *Techniques d'enquête*, 26, 113-122.
- Tillé, Y. (2002). Estimation sans biais par calage sur la répartition dans les plans simples sans remise. *Techniques d'enquête*, 28, 83-91.
- Tracy, D.S., Singh, S. et Arnab, R. (2003). Note sur le calage sous échantillonnage stratifié et double. *Techniques d'enquête*, 29, 111-116.
- Vanderhoef, C. (2001). Generalised calibration at Statistics Belgium. *Modèles et Applications* (Eds. J.J. Dreesbeke et L. Lebart), Paris : Dunod.
- Webber, M., Latouche, M. et Rancourt, E. (2000). Harmonised calibration of income statistics. Statistique Canada, document interne, avril 2000.
- Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937-951.
- Wu, C., et Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- Zheng, H., et Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.
- Zieschang, K.D. (1990). Sample weighting methods and estimation of totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.

- Estevao, V.M., et Särndal, C.-E. (2006). Survey estimates by calibration on complex auxiliary information. *Revue Internationale de Statistique*, 74, 127-147.
- Firth, D., et Benner, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, 60, 3-21.
- Folsom, R.E., et Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and poststratification. *Proceedings, Section on Survey Research Methods*, American Statistical Association, 598-603.
- Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Fuller, W.A., Longkhin, M.M., et Baker, H.D. (1994). Production de poids de régression en situation de non-réponse et application à la Natonwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.
- Hansen, M.H., et Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.
- Hartms, T. (2003). Extensions of the calibration approach: Calibration of distribution functions and its link to small area estimators. *Chinese document de travail numéro 13*, Federal Statistical Office, Germany.
- Hartms, T., et Duchesne, P. (2006). On calibration estimation for quantiles. *Techniques d'enquête*, 32, 37-52.
- Hidiroglou, M.A. (2001). L'échantillonnage double. *Techniques d'enquête*, 27, 157-169.
- Hidiroglou, M.A., et Särndal, C.-E. (1998). Emploi des données auxiliaires dans l'échantillonnage à deux phases. *Techniques d'enquête*, 24, 11-20.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- Huang, E.T., et Fuller, W.A. (1978). Nonnegative regression estimation for sample data. *Proceedings, Social Statistics Section*, American Statistical Association, 300-305.
- Isaki, C.T., Tsay, J.H., et Fuller, W.A. (2004). Pondération de données d'échantillon reposant sur des contrôles indépendants. *Techniques d'enquête*, 30, 39-49.
- Kalton, G., et Flores-Cerantes, I. (1998). Weighting methods. Dans *New Methods for Survey Research* (Eds. A.Westlake, J. Martin, M. Rigg et C. Skinner), Berkeley, U.K.: Association for Survey Computing.
- Kim, J., Breidt, F.J., et Opsomer, J.D. (2005). Nonparametric regression estimation of finite population totals under two-stage sampling. *Manuscript non publié*.
- Knothmerrus, P., et van Duin, C. (2006). Variances in repeated weighting with an application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, 565-584.
- Kott, P.S. (2004). Commentaire sur Demnat et Rao : Estimateurs de la variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 29-30.
- Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 149-160.
- Statistique Canada, N° 12-001-XPB au catalogue
- Ren, R. (2002). Estimation de la fonction de répartition et des fractiles d'une population finie. *Actes des journées de méthodologie statistique*, INSEE Méthodes, tome 1, 100, 263-289.
- Bureau of Latvia.
- Pihlkas, A. (2006). Non-linear calibration. *Proceedings, Workshop on Survey Sampling*, Ventspils, Latvia. Riga: Central Statistical Bureau of Latvia.
- 93-101.
- Park, M., et Fuller, W.A. (2005). Vers des poids de régression non généralisés pour les échantillons d'enquête. *Technique d'enquête*, 31, 93-101.
- Nieuwenbroek, N.J., Renssen, R.H. et Hofman, L. (2000). Towards a generalized weighting system. Dans *Proceedings, Second International Conference on Establishment Surveys*, American Statistical Association, Alexandria VA.
- Nieuwenbroek, N.J., et Boonstra, H.J. (2002). Bascula 4.0 for weighting sample survey data with estimation of variances. The Survey Statistician, Software Reviews, juillet 2002.
- Nieuwenbroek, N.J., et Boonstra, H.J. (2002). Bascula 4.0 for regression estimation. Central Bureau of Statistics, Pays-Bas.
- Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear class frequencies. *Statistics Finland Research Reports* 247.
- Myrskylä, M. (2007). Generalised regression estimation for domain calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.
- Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model-contribute, livre 2, 81-82.
- Montanari, G.E., et Ranalli, M.G. (2003). On calibration methods for design-based finite population inferences. *Bulletin of the International Statistical Institute*, 54^e session, volume LX, articles 191-202.
- Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.
- Montanari, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.
- Association, 99, 1131-1139.
- Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.
- Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.
- Lundström, S., et Särndal, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.
- Lemaître, G., et Durot, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- Lemaître, G., et Durot, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.
- Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7, 649-674.
- Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 37-49.
- Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 29, 37-49.
- Lehtonen, R., et Veijanen, A. (1998). Estimateur de régression généralisés logistiques. *Techniques d'enquête*, 24, 53-58.
- LeGuennec, J., et Sautory, O. (2002). CALMAR2 : une nouvelle version de la macro CALMAR de redressement d'échantillon par calage. *Actes des Journées de Méthodologie*, INSEE, Paris.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.
- Kravavickaitė, D., et Pihlkas, A. (2005). Estimation of a ratio in the finite population. *Informatica*, 16, 347-364.
- Kovacević, M.S. (1997). Calibration estimation of cumulative distribution and quantile functions from survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, 139-144.

Beaumont, J.-F. (2005a). Calibrated imputation in surveys under a quasi-modél-assisted approach. *Journal of the Royal Statistical Society B*, 67, 445-458.

Beaumont, J.-F. (2005b). L'utilisation de renseignements sur le processus de collecte des données pour traiter la non-réponse totale au moyen de l'ajustement de poids. *Techniques d'enquête*, 31, 249-254.

Beaumont, J.-F. (2004). Estimation robuste par la régression généralisée. *Techniques d'enquête*, 30, 217-231.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Bethlehem, J.G., et Keller, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28, 1026-1053.

Breidt, F.J., Claeskens, G. et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chambers, R.L., Dorfman, A.H. et Wehrly, T.E. (1993). Bias robust estimation in finite populations nonparametric calibration. *Journal of the American Statistical Association*, 88, 268-277.

Deming, W.E. (1943). *Statistical Adjustment of Data*. New York : John Wiley & Sons, Inc.

Dermati, A., et Rao, J.N.K. (2004). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête*, 30, 17-27.

Deville, J.-C. (1998). La correction de la nonréponse par calage par échantillonnage équilibré. Article présenté aux Congrès de l'ACFAS, Sherbrooke, Québec.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Document interne, I.N.S.E.E., Paris

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Duchesne, P. (1999). Estimateurs de calage robustes. *Techniques d'enquête*, 25, 47-60.

Dupont, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.

Ekholm, A., et Laaksonen, S. (1991). Weighing via response modelling in the Finnish Household Budget Survey. *Journal of Official Statistics*, 3, 325-337.

Estévez, V.M., et Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Estévez, V.M., et Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, 18, 233-255.

Estévez, V.M., et Särndal, C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics*, 20, 645-660.

Statistique Canada, N° 12-001-XPB an catalogue

II. Conclusion

Surdénonnement dans la base de sondage) et la non-réponse est exposé dans Särndal et Lundström (2005) et dans Kott (2006). Skinner (1998) discute de l'utilisation du calage en présence de non-réponse et d'erreurs de mesure. Son commentaire sur la nécessité de poursuivre les travaux de recherche en vue d'étudier les propriétés des estimations par calage en présence d'erreur non due à l'échantillonnage demeure un défi près de dix ans plus tard.

Si une question doit être choisie en vue de formuler la conclusion du présent exposé, il s'agit selon moi du concept d'information auxiliaire, qui est le concept central de l'article. Sans information auxiliaire, il n'y a pas de calage, car il n'existe rien sur quoi l'exécuter. J'ai mentionné par ailleurs que l'estimation par la régression est un autre moyen, qui suit un raisonnement différent, de tenir compte de l'information auxiliaire dans l'estimation.

L'un des objectifs du présent article était de broser le tableau de deux types de raisonnement et de souligner en quoi ils se distinguent. Des exemples montrent comment la réalisation d'un objectif d'estimation essentiellement semblable est abordée par certains auteurs suivant la logique du calage et par d'autres, suivant celle de la régression généralisée GREG (ou du moins principalement suivant l'un ou l'autre de ces types de raisonnement). Les estimateurs respectifs que ces auteurs finissent par recommander peuvent ou non donner des résultats concordants. Que les écarts aient ou non des conséquences importantes (pour ce qui est de la variance, du biais, de questions pratiques telles que la convergence et la transparence) dépend de la situation. Le présent article arrive peut-être à mieux faire comprendre ce qui distingue deux courants de pensée qui ont guidé les chercheurs spécialisés en échantillonnage.

Bibliographie

Alexander, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

Andersson, P.G., et Thorburn, D. (2005). Une distance de calage optimale menant à un estimateur par la régression optimale. *Techniques d'enquête*, 31, 103-107.

Ardilly, P. (2006). *Les techniques de sondage*. Paris : Éditions Technip.

Bankier, M.D., Rathwell, S. et Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 Canadian Census. Document de travail, Census Operations Section, Social Surveys Methods Division, Statistique Canada.

Bankier, M., Houle, A.M. et Luc, M. (1997). Calibration estimation in the 1991 and 1996 Canadian censuses. *Proceedings, Section on Survey Research Methods*, American Statistical Association, 66-75.

différentes) sont proposées dans Bethlehem (1988) et dans Füllmer, Loughin et Baker (1994). Nous pouvons aussi écrire

$$(\mathcal{D}^{\text{TV}}_k)^{-1} = \mathcal{D}(\theta^k M^k - 1)^{-1} \mathbf{x}^k \quad \text{ou} \quad M^k = (\mathcal{D} \mathbf{x}^k)^{-1} (\mathcal{D} \theta^k \mathbf{x}^k \mathbf{x}^k)^{-1} \mathbf{x}^k.$$

Dans la comparaison des \mathbf{x}^k , une référence commode est le « vecteur auxiliaire primitif », $\mathbf{x}^k = \mathbf{1}$ pour tout $k \in \mathcal{U}$, qui adonne $\mathbf{N}^{\text{AV}} = \mathbf{N}^{\text{AV}} = \mathbf{N}^{\text{AV}}/m$, $\mathbf{N}^{\text{AV}} = \mathbf{N}^{\text{AV}}/N$, le ratio

$$\mathcal{Y}^{\text{AV}}_{i,j} = \mathcal{D} \theta^k \mathbf{x}^k / \mathcal{D} \theta^k \quad \text{et} \quad \mathcal{Y}^{\text{AV}} = \mathcal{D} \theta^k / N.$$

Les ré pondants, avec biaisapprox $(\mathbf{N}^{\text{AV}})^{-1} = (\mathbf{N}^{\text{AV}})^{-1}$, où n est le nombre de

$$\text{biasrel}(\hat{Y}_{\text{CAL}}) = \frac{\text{biasapprox}(Y_{\text{CAL}})}{\sum_k^U (\theta^k M^k (I - I)^k)} = \frac{\text{biasapprox}(N_{\bar{Y}})}{N(\bar{Y}^{U\theta} - \bar{Y}^U)}$$

indique dans quelle mesure un vecteur candidat x_k réussit à contrôler le biais, comparativement au vecteur primitif. Nous recherchons un vecteur x_k qui donne un biais faible. Cependant, $\text{biaisrel}(x_k^{\text{cal}})$ n'est pas un indicateur de biais calculable, car il dépend de y_k non observé et de θ_k inobservable. Il nous faut un indicateur calculable qui s'approche de $\text{biaisrel}(x_k^{\text{cal}})$ et dépend du vecteur x_k , mais non des variables y_k qui peuvent être nombreuses dans

Il est facile de voir que $\text{baisrel}(X^{\text{CVAL}}) = 0$ dans le cas d'un vecteur x idéal (probablement inexistant) tel que $\phi_k = 1/\theta_k = \lambda_k x_k$ pour tout $k \in U$ et un vecteur constant λ_k .

Pour un vecteur x qui peut effectivement être construit dans le contexte de l'enquête, nous pouvons au moins obtenir les prédictions de ϕ^* . Si nous déterminons λ de manière à minimiser $\sum_{i=1}^n \theta^* \phi^*(x_i) - \lambda^* x_i^T$, nous obtenons $\lambda = \lambda^*$, où $\lambda^* = (\sum_{i=1}^n \theta^* x_i x_i^T)^{-1} (\sum_{i=1}^n \theta^* x_i)$; la valeur prévue de ϕ^* est $\phi^* = \lambda^* \phi^*$. Les premier et deuxième moments (pondérés par theta) des prédictions $M^* M^* \phi^* = N^{-1} \sum_{i=1}^n \theta^* \phi^* \phi^*^T = N^{-1} \sum_{i=1}^n \theta^* \phi^* \phi^*^T$ sont, respectivement, $M^* \sum_{i=1}^n \theta^* \phi^* \phi^*^T = N^{-1} \sum_{i=1}^n \theta^* \phi^* \phi^*^T$ et I / θ^* .

$$(\underline{n}_{\theta}/I - \underline{n}_W)(\underline{n}_{\theta}/I) = \tau^{(\theta; \underline{n}_W - \gamma_W)} \gamma_{\theta} \underline{n} \underline{\Sigma} \frac{\gamma_{\theta} \underline{n} \underline{\Sigma}}{I} = \bar{O}$$

où $M_U = \sum^U M^k/N$. Sørndal et Lundström (2007) montrent que, dans certaines conditions, la relation entre $\text{bias}(\text{rel}(Y^{\text{CAT}}))$ et Q est approximativement linéaire,

$$\frac{\partial \sigma_0}{\partial \sigma} - 1 \approx 1 - \text{biasrel}(Y_{\text{CAL}})$$

où $\phi = \sum_{j=1}^N \phi_j \mathbf{x}_j$ et $\phi_0 = (1/\theta)(\phi - 1/\theta)$ est la servit maximale de ϕ . Donc, si $\tilde{\phi}$ est calculable, il pourra servir d'indicateur pour comparer les différents vecteurs candidats. Nous obtenons un analogue calculable \tilde{Q} de Q en tant que variance des prédictions basses sur l'échantillon correspondantes $\mathbf{x}_j^* \mathbf{y}_j^*$ de sorte que

Sæmndal : La méthode de calage dans la théorie et la pratique des enquêtes

$$({}^{p^s}u - {}^{p^s}u) {}^{p^s}u = {}^c({}^{p^s}u - {}^qu) {}^qp^s\overline{\Sigma} \frac{{}^qp^s\overline{\Sigma}}{1} = \overline{Q}$$

no

$$\cdot \frac{\gamma p^s \underline{\Sigma}}{\gamma u \gamma p^s \underline{\Sigma}} = \overset{p^s}{\underline{u}} \cdot \frac{\gamma p^s \underline{\Sigma}}{\gamma p^s \underline{\Sigma}} = \frac{\gamma p^s \underline{\Sigma}}{\gamma u \gamma p^s \underline{\Sigma}} = \overset{p^s}{\underline{u}}$$

Nous nous attendons à ce que le biais relatif (*relbias*) diminue de manière approximativement linéaire à mesure que \hat{Q} augmente. Par conséquent, indépendamment des variables y , \hat{Q} peut être utilisé comme outil pour classer les divers vecteurs x en fonction de leur capacité à réduire le

Nous pouvons nous servir de \tilde{Q} comme outil de sélection des variables x qu'il convient d'inclure dans le vecteur x , par exemple par régression multiple ascendante de sorte que les variables soient ajoutées à x , une à la fois, celle qui entre à une étape donnée étant celle qui produit l'accroissement le plus important de \tilde{Q} . La méthode est décrite dans Sämndal et Lundström (2007).

10. Calage pour tenir compte d'autres erreurs non dues à l'échantillonnage

Les erreurs dues à la non-réponse jouent un rôle déterminant dans la qualité des statistiques publiées. En revanche, si nous examinons la place que pourrait tenir l'approche du calage dans le traitement des erreurs non dues à l'échantillonnage d'autres sources que la non-réponse, il n'est pas surprenant que la littérature soit jusqu'à présent moins abondante. Néanmoins, plusieurs auteurs esquisseront une approche de calage en vue d'intégrer également les erreurs de base de sondage, les erreurs de mesure et les valeurs aberrantes. Le calage pourrait offrir une théorie plus générale de l'estimation dans le contexte des enquêtes qui engloberaient les diverses erreurs non dues à l'échan-

Il comme ça.

Comme le souligne Deville (2004), le concept de calage s'applique aisément et efficacement à une grande variété de problèmes posés par les sondages. Selon lui, sa portée dépasse celle de l'estimation par la régression, une notion à laquelle certains semblent vouloir réduire l'approche du calage. Il expose sommairement comment cette approche permet de traiter plusieurs erreurs dues à la non-réponse.

Folsom et Singh (2000) présentent une méthode de calage des poids s'appuyant sur ce qu'ils appellent le modèle exponentiel généralisé (GEM). Elle comporte trois volets : le traitement des valeurs extrêmes, la correction de non-réponse et le calage par poststratification. Elle fournit des contrôles intégrés pour les valeurs extrêmes. Le calage pour corriger à la fois les erreurs de couverture (sous-

quadratique moyenne. Nous devons décider de limiter le biais autant que possible. L'approche du calage peut nous aider à construire un vecteur auxiliaire qui répond à cet objectif.

9.2 Calage pour la correction du biais de non-réponse

Faisant plus ou moins contraste avec la procédure classique, un certain nombre d'articles récents mettent l'accent sur l'approche du calage pour corriger la non-réponse. Certaines références récentes à cet égard sont Deville (1998, 2002), Ardlily (2006), chapitre 3, Skinner (1998), Folsom et Singh (2000), Fuller (2002), Lundström et Särndal (1999), Särndal et Lundström (2005), et Kott (2006).

L'approche du calage débute par l'évaluation de l'information auxiliaire totale disponible au niveau de l'échantillon (valeurs des variables auxiliaires observées pour les répondants et les non-répondants) et au niveau de la population (totaux auxiliaires connus de population).

L'objectif est de tirer le meilleur parti possible des deux sources combinées, afin de réduire le biais ainsi que la variance. Les poids de sondage sont modifiés, en une ou en deux étapes de calage, de façon qu'ils reflètent i) le résultat de la phase de réponses, ii) les caractéristiques individuelles des répondants et iii) l'information auxiliaire spécifiée. L'information peut se résumer comme il suit.

Niveau de la population : La valeur du vecteur auxiliaire \mathbf{x}_k^* est connue (spécifiée dans la base de sondage) pour chaque $k \in U$, donc est connue pour chaque $k \in s$ et chaque $k \in r$; $\sum_U \mathbf{x}_k^*$ est un total de population connu.

Niveau de l'échantillon : La valeur du vecteur auxiliaire \mathbf{x}_k^* est connue (observée) pour chaque $k \in s$, et est donc connue pour chaque $k \in r$; le total inconnu $\sum_U \mathbf{x}_k^*$ est estimé sans biais par $\sum_s d_k^* \mathbf{x}_k^*$.

Le calage sur cette information composite peut se faire en deux étapes (calcul de poids intermédiaires pour commencer, puis utilisation de ces poids à la deuxième étape pour produire les poids finaux) ou directement en une seule étape. En principe, les différences de biais et de variance des estimations devraient être modestes. Dans l'option en une seule étape, le vecteur auxiliaire combiné et l'information correspondante sont

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^0 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_k^* \\ \sum_U \mathbf{x}_k^* - \sum_s d_k^* \mathbf{x}_k^* \end{pmatrix}.$$

En utilisant une extension de la méthode du vecteur instrumental décrite à la section 4.3, nous recherchons les poids calés $w_k = d_k^* v_k$, où $v_k = F(\lambda^* z_k)$ est le facteur de correction de la non-réponse, avec un vecteur λ^* déterminé

par l'équation de calage $\sum_r w_k \mathbf{x}_k = \mathbf{X}$; l'estimateur par calage résultant est $\hat{Y}_{CAL} = \sum_r w_k y_k$. Il suffit de spécifier la valeur du vecteur instrumental \mathbf{z}_k pour les répondants; \mathbf{z}_k peut être différent de \mathbf{x}_k^* . La fonction $F(\cdot)$ joue le même rôle qu'aux section 4.2 et 4.3. Ici, $F(\lambda^* z_k)$ estime implicitement la probabilité de réponse inverse, $\phi_k = 1/\theta_k$ comme l'ont souligné Deville (2002), Dupont (1995) et Kott (2006). Dans le cas linéaire, $F(u) = 1 + u$ et $v_k = 1 + \lambda^* z_k$, avec $\lambda^* = (\sum_U \mathbf{x}_k^* - \sum_r d_k^* \mathbf{x}_k^*) (\sum_r d_k^* \mathbf{z}_k \mathbf{x}_k^*)^{-1}$. Bien qu'elles ne soient observées que pour les éléments échantillonnés, les variables qui constituent le vecteur \mathbf{x}_k^* peuvent être de la plus haute importance pour la réduction du biais de non-réponse (quoique moins importantes que les \mathbf{x}_k^* pour la réduction de la variance). Par exemple, Beaumont (2005b) discute des variables du processus de collecte des données qui peuvent être utilisées pour construire la composante vectorielle \mathbf{x}_k^* .

9.3 Construction du vecteur auxiliaire

Dans certaines enquêtes, les variables auxiliaires possibles abondent, comme le signalent, par exemple, Rizzo, Kallon et Brick (1996), de même que Särndal et Lundström (2005). Ainsi, pour les enquêtes auprès des ménages et des particuliers en Scandinavie, on peut obtenir une foule de variables auxiliaires éventuelles par appartenance des données d'enquête à celles des registres administratifs de haute qualité existants. Il convient ensuite de décider lesquelles de ces variables devraient être incluses dans le vecteur auxiliaire \mathbf{x}_k^* afin que celui-ci soit aussi efficace que possible, notamment en ce qui concerne la réduction du biais. Comme le mentionnent Rizzo, Kallon et Brick (1996), le choix des variables auxiliaires est probablement plus important que celui de la méthode de pondération.

Examinons le biais quand $\mathbf{z}_k = \mathbf{x}_k$. Nous devons comparer divers vecteurs \mathbf{x}_k^* afin de choisir, en dernière analyse, celui qui produira vraisemblablement le biais le plus faible. (Nous posons que \mathbf{x}_k^* est tel que $\mathbf{1}' \mathbf{x}_k^* = 1$ pour tout k et un vecteur constant $\mathbf{1}$, ce qui est le cas de nombreux vecteurs \mathbf{x}_k^* , y compris les exemples 1 à 5 présentés au début de la section 2.) Nous obtenons une bonne approximation du biais de \hat{Y}_{CAL} par linéarisation de Taylor de la forme biaisapprox(\hat{Y}_{CAL}) = $(\sum_U \mathbf{x}_k^*)' (\mathbf{B}_{U;0} - \mathbf{B}_U)$, où intervient la différence entre le coefficient de régression pondéré $\mathbf{B}_{U;0} = (\sum_U \theta_k \mathbf{x}_k^* \mathbf{x}_k^*)^{-1} \sum_U \theta_k \mathbf{x}_k^* y_k$ et le coefficient non pondéré $\mathbf{B}_U = (\sum_U \mathbf{x}_k^* \mathbf{x}_k^*)^{-1} \sum_U \mathbf{x}_k^* y_k$. À moins que tous les θ_k soient égaux, le biais causé par l'écart entre les deux vecteurs de coefficients de régression pourrait être important, même si \mathbf{x}_k^* semble être un « bon vecteur auxiliaire ». Cette expression du biais approximatif (nearbias) est donnée dans Särndal et Lundström (2005). Des expressions apparentées du biais (sous des conditions

L'information utilisée à cette étape provient souvent du regroupement des éléments échantillonnés. Enfin, si des totaux de population fiables sont disponibles, les poids de sondage corrigés sont calés sur ces totaux.

Au Canada, la méthodologie de l'Enquête sur la population active, décrite dans Statistique Canada (1998), est un exemple de cette pratique répandue. Un poids de sondage (modifié) est d'abord calculé pour un ménage donné, par multiplication de trois facteurs. Le produit du poids de sondage et d'un facteur de correction de la non-réponse est appelé le sous-poids. À la dernière étape, les sous-poids sont calés sur des estimations de population postcensitaire très précises par groupe d'âge, sexe et région intraprovinciale. Les poids finaux ont les propriétés souhaitées de convergence, dans les régions d'une province, avec les estimations postcensitaires. Le biais de non-réponse, qui persiste dans les estimations résultantes est inconnu, mais considéré comme modeste.

La méthode classique est intégrée dans le type d'estimateur $\hat{Y} = \sum_r d_k^k (1/\theta_k^k) y_k^k$, où θ_k^k a été estimé par θ_k^k à une étape préliminaire, en modélisant la réponse (c'est-à-dire la proposition à répondre). Ce que demande la théorie au statisticien, à savoir formuler le « modèle de réponse réelle », capable de fournir des valeurs θ_k^k exactes, sans biais, n'est pas une tâche facile. Toutefois, dans de nombreuses enquêtes, les facteurs $1/\theta_k^k$ sont appliqués machinalement, sans esprit critique, par exemple par extension directe dans les strates déjà utilisées pour la sélection de l'échantillon.

La méthode classique est appliquée, par exemple, dans Ekholm et Laaksonen (1991) et dans Rizzo, Kalton et Brick (1996).

Souvent, les praticiens agissent comme si l'estimateur résultant $\hat{Y} = \sum_r d_k^k (1/\theta_k^k) y_k^k$ (découlant d'une modélisation plus ou moins pénétrante de la réponse pour essayer d'obtenir les θ_k^k) était essentiellement sans biais, ce qu'il n'est pas (à moins qu'on ait eu la chance de spécifier le modèle idéal). Ils se comportent (pour les besoins de l'estimation de la variance, par exemple) comme si π_k^k était la probabilité de sélection réelle de l'élément k dans une étape unique de sélection, ce qui n'est définitivement pas le cas. Cette pratique, dont l'origine remonte à un passé idyllique, devient de plus en plus contestable à mesure que grimpent subrepticement les taux de non-réponse. Le remplacement de θ_k^k par θ_k^k donne inévitablement lieu à un biais. Il y a plusieurs décennies, les taux de non-réponse n'étaient habituellement que de quelques pour cents et il était justifiable d'ignorer ce biais, mais leur croissance galopante rend cette pratique indéfendable aujourd'hui. Selon les premiers principes, l'objectif est de produire une estimation sans biais et non une estimation où le carré du biais est un contributeur dominant (et inconnu) à l'erreur

9. Calage pour corriger la non-réponse

9.1 Correction classique de la non-réponse

Les auteurs d'articles plus récents reviennent sur la question de la double pondération ménage et personne. Certains adoptent l'approche du calage et d'autres, celle de la régression généralisée GREG, Tsay et Fuller (2004) en font un problème de pondération par calage; ils appliquent des poids calés aux totaux de contrôle au niveau du ménage, ainsi qu'au niveau de la personne et ne fournissent aucun modèle auxiliaire explicite. Par contre, Steel et Clark (2007) suivent l'approche GREG, avec spécification de modèles auxiliaires linéaires et des structures de variance connexes.

Nombre de bons articles théoriques ont pour contexte le cas simple de la section 2, qui inclut l'absence totale de non-réponse. Ce sont de bonnes théories, mais applicables dans des conditions qui ne se réalisent que rarement, voire jamais. (À titre d'auteur d'articles dans ce domaine, je ne suis moi-même pas irréprochable.) La non-réponse existe dans pratiquement toutes les enquêtes. Bien qu'elle ne soit pas souhaitable, il s'agit d'un phénomène naturel dont la théorie devrait tenir compte d'emblée, en adoptant une perspective de sélection à deux phases. Dans de nombreuses enquêtes, les taux de non-réponse sont aujourd'hui très élevés si on les compare à ceux d'il y a 40 ans, qui étaient si faibles qu'on pouvait essentiellement ne pas en tenir compte. De nos jours, la théorie de l'échantillonnage doit s'attacher de plus en plus souvent aux conséquences indéfinissables de la non-réponse. En particulier, un objectif immédiat est d'examiner le biais et d'essayer de le réduire autant que possible.

Soit un échantillon probabiliste s tiré de $U = \{1, 2, \dots, k, \dots, N\}$. Il y a non-réponse, ce qui donne un ensemble de réponses r_i qui est un sous-ensemble de s ; la valeur de la variable étudiée y_k est observée pour $k \in r$ uniquement. La probabilité de réponse inconnue de l'élément k est $\Pr(k \in r|s) = \theta_k^k$. Nous écartons l'estimateur sans biais $\hat{Y} = \sum_r d_k^k \phi_k^k y_k^k$, parce que $\phi_k^k = 1/\theta_k^k$ est inconnue. Si nous voulons retenir la notion de somme totale, ou non-réponse d'une unité, en se fondant sur la « modélisation de la non-réponse » se fait depuis longtemps. Le calage offre une nouvelle perspective.

Dans ce que nous pourrions appeler la « méthode classique », les poids de sondage probabilistes $d_k^k = 1/\pi_k^k$ sont d'abord corrigés de la non-réponse et, éventuellement, d'autres imperfections, telles que les valeurs aberrantes.

éléments sous-échantillonnés dans les grappes sélectionnées au deuxième degré) et l'échantillonnage à deux phases tient au fait que l'information totale peut comporter plus d'une composante. Il peut exister (a) de l'information au niveau de la grappe (au sujet des grappes), (b) de l'information au niveau de l'élément pour toutes les grappes et (c) de l'information au niveau de l'élément pour les grappes sélectionnées uniquement. Ici encore, les auteurs appartiennent à deux écoles, certains exploitant l'information par l'approche du calage et les autres choisissant la route de la régression généralisée GREG.

Estévez et Sæmdal (2006) conçoivent l'estimation par calage sous échantillonnage à deux degrés classique, où l'information composite est spécifiée comme il suit : i) pour la population de grappes U , il existe un total commun $\sum_{i \in U} \mathbf{x}^{(c)i}$, où $\mathbf{x}^{(c)i}$ est une valeur du vecteur auxiliaire associée à la grappe U , pour $i \in U$; ii) pour la population d'éléments $U = \bigcup_{i \in U} U$, il existe un total commun $\sum_{i \in U} \mathbf{x}^k$, où la valeur du vecteur auxiliaire \mathbf{x}^k est associée à l'élément $k \in U$. Supposons que l'on doit produire à la fois les statistiques de grappe et les statistiques d'élément dans une enquête, si bien qu'il faut estimer le total de population $Y = \sum_{i \in U} Y^{(c)i}$ et le total de population d'éléments $X = \sum_{i \in U} X^k$.

Si aucune relation n'est imposée entre les poids des grappes w_U et les poids des éléments w_k , les premiers sont calés de manière à satisfaire $\sum_{i \in U} w_U \mathbf{x}^{(c)i} = \sum_{i \in U} \mathbf{x}^{(c)i}$ et les seconds, de manière à satisfaire $\sum_{k \in U} w_k \mathbf{x}^k = \sum_{k \in U} \mathbf{x}^k$. (Ici, s_1 est l'échantillon de grappes tiré de U ; s_2 est l'échantillon d'éléments tiré de U , et $s = \bigcup_{i \in s_1} s_2$.) Alors, $Y_{\text{CAL}} = \sum_{i \in s} w_U Y^{(c)i}$ estime le total de population de grappes Y et $X_{\text{CAL}} = \sum_{k \in s} w_k X^k$ estime le total de population d'éléments X .

Dans la pratique, on recourt souvent à la pondération intégrée, qui consiste à imposer une relation commune entre les poids des grappes w_U et les poids w_k des éléments compris dans les grappes sélectionnées. Estévez et Sæmdal (2006) examinent deux formes de pondération intégrée.

L'une d'elles consiste à imposer $w_k = d_{k|U} w_U$, où $d_{k|U}$ est l'inverse de la probabilité d'inclusion de l'élément k dans la grappe i . (Par exemple, dans l'échantillonnage en grappes à grappe i , quand tous les éléments k d'une grappe i ont le même poids pour le calcul des statistiques d'élément, et ce même poids est également utilisé pour calculer les statistiques de grappe.) L'équation de calage $\sum_{k \in U} w_k \mathbf{x}^k = \sum_{k \in U} \mathbf{x}^k$ se lit alors $\sum_{i \in U} w_U \sum_{k \in i} d_{k|U} \mathbf{x}^k = \sum_{k \in U} \mathbf{x}^k$. Les poids de grappe w_U sont maintenant calculés en minimisant $\sum_{i \in U} (w_U - d_{k|U})^2 / d_{k|U}$ sous la contrainte de l'équation de calage qui tient compte des deux types d'information :

8.3 Pondération des ménages et pondération des personnes

Dans certaines grandes enquêtes sociales, l'objectif est de produire des estimations au niveau du ménage ainsi qu'au niveau de la personne, si bien que certaines variables étudiées sont des variables du ménage (grappe) et d'autres, certaines variables de la personne (élément). Par conséquent, un certain nombre d'auteurs ont étudié la situation de l'échantillonnage en grappes à un degré ($d_{k|U} = 1$) et de la pondération intégrée qui consiste à attribuer le même poids à tous les membres d'un ménage sélectionné, poids qui est également utilisé pour produire les statistiques au niveau du ménage. Une solution générale de ce problème de pondération, si l'information au niveau du ménage et l'information au niveau de la personne sont l'une et l'autre spécifiées, consiste à obtenir les poids des ménages w_U , calés comme dans l'équation (8.1) avec $d_{k|U} = 1$, puis à prendre $w_k = w_U$.

Dans plusieurs articles, l'accent est mis sur les valeurs du vecteur auxiliaire \mathbf{x}^k attribuées aux personnes. Alexander (1987) calcule des poids qui minimisent la distance du chi-carré, tandis que Lemaître et Dufour (1987) et Nieuwenbroek (1993) calculent les poids intégrés à l'aide d'un estimateur GREG. La méthode de Lemaître et Dufour procède par construction indirecte d'une « valeur de vecteur auxiliaire équilibrée » s'appliquant à toutes les personnes membres d'un ménage sélectionné. Leur résultat est calculable par la méthode directe exposée à la section 8.2.

Une fois les poids de grappe w_U déterminés, le calcul des poids d'élément $w_k = d_{k|U} w_U$ s'ensuit.

Une autre méthode de pondération intégrée raisonnable consiste à imposer que $\sum_{k \in U} w_k = N^i w_U$. Par exemple, pour l'échantillonnage en grappes à un degré, cela implique que le poids de grappe w_U est égal à la moyenne des poids d'élément w_k contenus dans la grappe.

L'échantillonnage à deux degrés est également traité dans Kim, Breidt et Opsomer (2006). Ces auteurs supposent que des données auxiliaires existent pour les grappes, par la voie d'une variable de grappe quantitative unique $X^{(c)i}$, mais non pour les éléments. Ils développent et étudient un estimateur de type GREG du total de population d'éléments $X = \sum_{i \in U} X^k$, $Y = \sum_{i \in U} Y^i + \sum_{i \in s_1} d_{i|U} Y^i$, où Y^i est sans biais sous le plan pour le total de population de grappes $Y = \sum_{i \in U} Y^i$ et Y^i est obtenu par un ajustement par régression polynomiale locale. L'estimateur peut être exprimé sous la forme linéairement pondérée en utilisant des poids qui sont calés sur les totaux de population des puissances de la variable de grappe $X^{(c)i}$.

$$(8.1) \quad \left(\sum_{i \in U} \mathbf{x}^{(c)i} \right) = \left(\sum_{i \in s_1} w_U \sum_{k \in i} d_{k|U} \mathbf{x}^k \right)$$

plusieurs formes d'chantillonnage double. Dans le cas hirarchique (ou *embote*) (chantillonnage  deux phases classique), l'chantillon de premire phase s_1 est tir de U_1 et l'chantillon de seconde phase s_2 est un sous-chantillon tir de s_1 , de sorte que $U \subset s_1 \subset s_2$. Nous pouvons distinguer deux cas non hirarchiques (ou non *embotes*). Dans le premier, s_1 est tir de la base de sondage U_1 et s_2 de la base de sondage U_2 , o U_1 et U_2 couvrent la mme population U , et les units d'chantillonnage peuvent tre dfinies diffremment dans les deux bases. Dans le deuxime cas non hirarchique, s_1 et s_2 sont tirs indpendamment de U .

Afin d'illustrer comment l'information composite intervient dans l'estimation, considrons le cas hirarchique. Les poids de sondage sont $d_k = 1/\pi_k$ (s_1 chantillonn dans U); $d_k = 1/\pi_k(\pi_k = \pi_{k|s_1})$ dans le sous-chantillonnage de s_2  partir de s_1). Les poids de sondage combins est $d_k = d_k d_k$. L'estimateur sans biais lmentaire $\hat{Y} = \sum_{s_2} d_k \hat{y}_k$ peut tre amlior grce  l'utilisation d'information auxiliaire, spcifie ici  deux niveaux.

Niveau de la population : La valeur du vecteur \mathbf{x}_k est connue (donne dans la base de sondage) pour chaque $k \in U$, de sorte qu'elle est connue pour chaque $k \in s_1$ et pour chaque $k \in s_2$; $\sum_{s_2} \mathbf{x}_k$ est un total vectoriel de population connu.

Niveau du premier chantillon : La valeur du vecteur auxiliaire \mathbf{x}_{2k} est connue (observe) pour chaque $k \in s_1$ et, par consquent, pour chaque $k \in s_2$; le total inconnu $\sum_{s_2} \mathbf{x}_{2k}$ est estim sans biais par $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$.

Quel est le meilleur moyen de tenir compte de cette information composite? Dans une adaptation de l'approche GREG, Sndal et Swensson (1987) ont formul deux modles auxiliaires finitaires, le premier spcifi en fonction du vecteur \mathbf{x}_{1k} et le second tenant compte aussi du vecteur \mathbf{x}_{2k} . Les deux modles sont ajusts et les prdictions rsultantes, de deux types, sont utilises pour crer un estimateur GREG appropri $\hat{Y}_{\text{GREG}}^Y = \sum_{s_2} \mathbf{y}_k$. Dupont (1995) fait le commentaire important que l'information composite donne appelle « deux approches naturelles diffrentes ». Outre l'approche GREG, il existe une approche par calage qui donnera les poids finaux w_k pour un estimateur par calage $\hat{Y}_{\text{CAL}} = \sum_{s_2} w_k \mathbf{y}_k$. Il est intressant de comparer les rsultats de ces deux approches. L'une et l'autre offrent plus d'une option. Dans l'approche GREG, il existe divers moyens de formuler les modles auxiliaires finitaires et leur structure de variance respective. Dans l'approche par le calage, diverses formulations des quations de calage sont possibles.

Voici, par exemple, une option de calage en deux tapes : d'abord trouver les poids intermdiaires w_{1k} qui

$$\text{o } \mathbf{x}_k \text{ est le vecteur auxiliaire combin} \\ \sum_{s_2} w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k \left(\sum_{s_2} \mathbf{x}_{2k} \right) = \sum_{s_1} \mathbf{x}_{1k} \left(\sum_{s_2} d_{1k} \mathbf{x}_{2k} \right).$$

satisfont

Alternativement, dans une option  une seule tape, nous dterminons les poids w_k directement pour satisfaire

En gnral, les poids finaux w_k ne sont pas identiques dans les deux options. Supposons que $\sum_{s_2} \mathbf{x}_{1k}$ est un total \mathbf{x}_1 import. Si nous examinons la situation de plus prs, nous constatons que l'option en deux tapes ncessite plus d'information, parce que les valeurs connues \mathbf{x}_{1k} sont requises individuellement pour $k \in s_1$, tandis que dans l'option  une seule tape, il suffit qu'elle soit disponible pour $k \in s_2$. Nous pourrions donc nous attendre  un certain avantage de l'option  deux tapes en ce qui concerne la variance, puisque $\sum_{s_2} w_{1k} \mathbf{x}_{2k}$ est souvent plus exact (en tant qu'estimateur de $\sum_{s_2} \mathbf{x}_{2k}$) que $\sum_{s_1} d_{1k} \mathbf{x}_{2k}$ dans la procdure  une seule tape. Nanmoins, cette attente n'est pas toujours confirme et la mthode  une seule tape peut tre suprieure, par exemple quand \mathbf{x}_1 et \mathbf{x}_2 sont faiblement corrls.

Dupont (1995), ainsi qu'Hydroglogou et Sndal (1998) examinent les liens qui existent, sans surprises, entre les deux approches. Un estimateur GREG, driv de modles adjoints ayant des structures de variance spcifiques, pourrait tre identique  l'estimateur par calage si les poids de ce dernier sont cals d'une certaine faon. Dans d'autres cas, les diffrences pourraient tre faibles.

L'efficacit de diverses options dpend, de manire assez subtile, de la configuration des corrlations entre \mathbf{y}_k , \mathbf{x}_{1k} et \mathbf{x}_{2k} . Par exemple, dans quelle mesure \mathbf{x}_1 et \mathbf{x}_2 sont-ils complmentaires, dans quelle mesure sont-ils des substituts l'un de l'autre? Dans l'approche GREG, il est difficile, voire mme futile, de dfinir avec prcision une structure de variance qui reflte vraiment la « ralit » qui sous-tend les donnes. L'approche du calage est plus directe et Estvao et Sndal (2002, 2006) explorent certaines de ses possibilits.

8.2 Information composite dans les plans d'chantillonnage  deux degrs

Le point commun entre l'chantillonnage  deux degrs classique (grappes chantillonnes au premier degr,

pour penser que l'estimateur par calage se compare favorablement (variance plus faible, tout en maintenant la quasi-absence de biais sous le plan)    d'autres estimateurs,   tablissant d'autres arguments que le calage, tout en s'appuyant sur la m  me information auxiliaire.

7. Comparaison du calage    d'autres approches

Comme beaucoup l'ont soulign  , les utilisateurs voient dans le calage un moyen simple et convaincant d'int  grer l'information auxiliaire, pour des param  tres simples (section 4), ainsi que pour des param  tres plus complexes, tels que les quantiles, les ratios et d'autres (section 6). Sa simplicit   et son caract  re pratique sont des avantages ind  niables, mais peut-on aussi affirmer que le calage est   sup  rieur du point de vue th  orique  ? Existe-t-il des situations o   l'on peut montrer que le calage donne des r  ponses plus exactes et/ou plus satisfaisantes    des questions d'importance que d'autres approches fond  es sur le plan de sondage?

La section 4.5 donne une raison de penser que l'approche du calage pourrait   tre sup  rieure    l'approche GREG, parce que le calage bas   sur un mod  le pourrait donner des estimations plus pr  cises que l'estimateur GREG non lin  aire, pour le m  me mod  le auxiliaire. La section 7.1 qui suit offre un autre exemple o   l'approche du calage et l'approche GREG produisent des r  ponses divergentes, l'avantage allant    la m  thode de calage.

7.1 Un exemple d'estimation par domaine

L'exemple pr  sent   ici, tir   d'Est  vao et S  mard (2004), illustre, pour une situation pratique simple, un conflit entre les r  sultats de l'approche GREG et de celle du calage. Le contexte est l'estimation du total y d'une sous-population (domaine).

Un   chantillon probabiliste s est tir   de $\{1, 2, \dots, k, \dots, N\}$; les poids de sondage connus sont $d_k = 1/\pi_k$. Soit U_a un domaine; $U_a \subset U$. L'indicateur de domaine est δ_{ak} , dont la valeur est $\delta_{ak} = 1$ si $k \in U_a$ et $\delta_{ak} = 0$ autrement. La grandeur estim  e est le total de domaine $X_a = \sum_U \delta_{ak} x_k$, o   $y_{ak} = \delta_{ak} y_k$, et y_k est observ   pour $k \in s$. L'estimateur $\hat{X}_{aHT} = \sum_s d_k y_{ak}$ de Horvitz-Thompson est sans biais sous le plan de sondage, mais sa pr  cision est faible, surtout si le domaine est petit, et l'utilisation d'information auxiliaire permettra de l'am  liorer. Nous sp  cifions une valeur du vecteur auxiliaire \mathbf{x}_k pour chaque $k \in U$.

Comme il est fr  quent en pratique, les   l  ments qui appartiennent    un domaine d'int  r  t ne sont pas pr  cis  s dans la base de sondage (s'ils le sont, on dispose d'information tr  s puissante d  s le d  part, mais souvent, les conditions r  elles ne sont pas aussi favorables). N  anmoins,

L'estimation des quantiles illustre bien le fait que la m  thode de calage peut   tre ex  cut  e de plus d'une fa  on lorsqu'on estime des param  tres un peu plus complexes. Les deux m  thodes mentionn  es ici donnent une estimation presque sans biais sous le plan. Les poids de Harms et Duchesne (2006) sont polyvalents, ind  pendants de la variable y . Par contre, la m  thode de Rueda et coll. (2007) n  cessite un nouvel ensemble de poids pour chaque nouvelle variable y . Des donn  es empiriques, obtenues par simulation, donnent    penser que les deux m  thodes se comparent favorablement aux m  thodes ant  rieures d'estimation des quantiles, non fond  es explicitement sur l'approche du calage (mais sur la m  me information auxiliaire).

L'extension de la m  thode du calage    l'estimation d'autres param  tres complexes, tels que le coefficient de Gini, est esquis  e dans Harms et Duchesne (2006).

6.2 Calage pour d'autres param  tres complexes

Plikusas (2006), ainsi que Kravavickait   et Plikusas (2005) examinent l'estimation par calage de certaines fonctions de totaux de population. Leur expression   calage non lin  aire   signifie    fonction non lin  aire de totaux    et n'est pas utilis  e ici). Un exemple simple est l'estimation du ratio de deux totaux, $R = \sum_U y_{1k} / \sum_U y_{2k}$, o   y_{1k} et y_{2k} sont les valeurs, pour l'  l  ment k , des variables y_1 et y_2 , respectivement. (En fait, la fonction de r  partition (6.1) est   galement du type ratio, avec $y_{2k} = 1$ et $N = \sum_U 1$ comme total au d  nominateur). Ces auteurs   tudient l'estimateur par calage $R_{CAL} = \sum_s w_k y_{1k} / \sum_s w_k y_{2k}$. Les poids w_k , communs au num  rateur et au d  nominateur, sont d  termin  s par l'information auxiliaire   nonc  e comme il suit : il existe une variable auxiliaire, x_{1k} , pour y_{1k} , et une autre, x_{2k} , pour y_{2k} ; le ratio des totaux $R_0 = \sum_U x_{1k} / \sum_U x_{2k}$ est une valeur connue, obtenue par un d  nombrement complet ant  rieur ou provenant d'une autre source fiable. L'  quation de calage propos  e est $\sum_s w_k e_k = 0$, o   $e_k = x_{1k} - R_0 x_{2k}$. Parce que $\sum_U e_k = 0$, selon la m  thode de la distance du chi-carr   minimale, les poids sont

$$w_k = d_k \left\{ 1 - \left(\sum_s d_k e_k \right) \left(\sum_s d_k e_k^2 \right)^{-1} e_k \right\}.$$

Ces poids extraient correctement la valeur connue du ratio R_0 ; en posant que $y_{1k} = x_{1k}$ et $y_{2k} = x_{2k}$ dans R_{CAL} , nous obtenons

$$\sum_s w_k x_{1k} - R_0 \sum_s w_k x_{2k} = 0.$$

Les donn  es empiriques pr  sent  es dans Plikusas (2006), ainsi que dans Kravavickait   et Plikusas (2005) donnent   

6.1 Calage de l'estimation des quantiles

La médiane et d'autres quantiles de la population finie dans le cas des enquêtes économiques. Afin d'estimer les quantiles, il faut d'abord estimer la fonction de répartition

répandue, plusieurs auteurs ont considéré l'estimation des quantiles, avec ou sans l'utilisation d'information auxiliaire. Les auteurs d'articles plus récents se sont tournés vers la méthode de calage dans le même but, dont Kovachévitch (1997), Wu et Sitter (2001), Ren (2002), Tillé (2002), Harms et Duchesne (2006), et Rueda et coll. (2007). Comme l'illustrent ces articles, il existe plus d'un moyen d'appliquer cette méthode. Le caractère non lisse de la fonction de répartition de la population finie cause certaines complexités dont la résolution varie selon les auteurs.

Soit $\Delta(\cdot)$ la fonction de Heaviside, définie pour tout réel z de manière que $\Delta(z) = 1$ si $z \geq 0$ et $\Delta(z) = 0$ si $z < 0$. La fonction de répartition inconnue de la variable étudiée y est

$$F_y(t) = \frac{1}{N} \sum_{j=1}^N \Delta(t - y_j). \quad (6.1)$$

Le quantile α de la population finie est défini comme étant $\bar{Q}_{y,\alpha} = \inf\{t | F_y(t) \geq \alpha\}$. La variable auxiliaire x_j , possédant les valeurs x_j , possède la fonction de répartition $F_x(t) = (1/N) \sum_{j=1}^N \Delta(t - x_j)$ avec le quantile α dénoté $\bar{Q}_{x,\alpha}$, $j = 1, 2, \dots, J$. Un estimateur naturel de $F_y(t)$ basé sur les poids de sondage $d_k = 1/\pi_k$ est

$$\hat{F}_y(t) = \frac{1}{I} \sum_{j=1}^I d_j \Delta(t - y_j).$$

Un estimateur par calage de $F_y(t)$ prend la forme

$$\hat{F}_{y,\text{CAL}}(t) = \frac{1}{I} \sum_{j=1}^I w_k \Delta(t - y_j). \quad (6.2)$$

où les poids w_k sont calculés comme il convient sur une information auxiliaire spécifiée. Puis, nous tirons de $\hat{F}_{y,\text{CAL}}(t)$ l'estimateur du quantile α donné par $\hat{Q}_{y,\alpha} = \inf\{t | \hat{F}_{y,\text{CAL}}(t) \geq \alpha\}$. Une formule analogue à (6.2) est vérifiée pour $\hat{F}_{x,\text{CAL}}(t)$.

Sans référence explicite à un modèle, Harms et Duchesne (2006) spécifient l'information disponible pour le calage comme étant une taille de population connue, N_j , et les quantiles de population connue $\bar{Q}_{x,\alpha}$ pour $j = 1, 2, \dots, J$. L'information auxiliaire complète, avec les valeurs $\mathbf{x}_k = (x_{k1}, \dots, x_{kJ})'$ connues pour $k \in U$, n'est pas nécessaire. (Toutefois, dans la pratique, l'information complète serait généralement requise, parce qu'il est peu probable que les valeurs exactes des quantiles de plusieurs

variables x puissent être importées des sources externes.) Ils déterminent les poids w_k de façon à minimiser la distance du chi-carré $\sum_s (w_k - d_k)^2 / 2d_k q_k$, pour la valeur spécifiée de q_k , sous la contrainte des équations de calage

$$\sum_{j=1}^J w_k \cdot \bar{Q}_{x,\text{CAL},\alpha} = N_j = \bar{Q}_{x,\alpha}, \quad j = 1, 2, \dots, J$$

pour des estimations définies convenablement $\bar{Q}_{x,\text{CAL},\alpha}$. Maintenant, si nous décidons de spécifier $\bar{Q}_{x,\text{CAL},\alpha}$ comme $\bar{Q}_{x,\alpha} = \inf\{t | \hat{F}_{x,\text{CAL}}(t) \geq \alpha\}$, il serait généralement impossible de trouver une solution exacte au problème de calage tel qu'il est énoncé. Harms et Duchesne choisissent plutôt de lui substituer des estimateurs lissés, qu'ils appellent « estimateurs interpolés de répartition » des fonctions de répartition $F_x(t)$, $j = 1, 2, \dots, J$. Ils remplacent $\Delta(\cdot)$ par une fonction légèrement modifiée. Il est alors possible d'obtenir les poids w_k , ainsi qu'une fonction de répartition estimée correspondante $\hat{F}_{y,\text{CAL}}(t)$; enfin, ils estiment $\bar{Q}_{y,\alpha}$ comme $\bar{Q}_{y,\alpha} = \hat{F}_{y,\text{CAL}}^{-1}(\alpha)$.

Les poids calculés résultants w_k nous permettent d'extraire les quantiles de population connus des variables auxiliaires. La chose est rassurante, car on s'attendrait à ce que ces poids produisent des estimateurs raisonnables des quantiles de la variable étudiée y . De surcroît, dans le cas d'une variable auxiliaire scalaire unique x , l'estimateur par calage résultant donne des quantiles de population exacts pour y quand la relation entre y et x est parfaitement linéaire, c'est-à-dire quand $y_k = \beta x_k$ pour tout $k \in U$. Une idée faisant intervenir des fonctions de répartition lissées est également mentionnée dans Tillé (2002).

La méthode mathématiquement plus simple de Rueda et coll. (2007) est une application du calage fondée sur un modèle en ce que leur calage se fait par rapport à un total de population des valeurs *prévues* de y . Elle requiert une information auxiliaire complète. Partant de la valeur connue \mathbf{x}_k , elle consiste à calculer d'abord les prédictions linéaires $\hat{y}_k = \beta' \mathbf{x}_k$ pour $k \in U$, avec $\beta = (\sum_s d_s q_s \mathbf{x}_s' \mathbf{x}_s')^{-1} (\sum_s d_s q_s \mathbf{x}_s' y_s)$, où $d_k = 1/\pi_k$ et les q_k sont des facteurs d'échelle spécifiés. Les poids w_k sont obtenus par minimisation de la distance du chi-carré sous la contrainte des équations de calage énoncées en fonction des prédictions, de façon à réaliser la convergence à J points choisis arbitrairement t_j , $j = 1, \dots, J$:

$$\frac{1}{I} \sum_{j=1}^J w_k \Delta(t_j - y_k) = F_y(t_j), \quad j = 1, \dots, J$$

où $F_y(t_j)$ est la fonction de répartition en population finie des prédictions y_k , évaluées à t_j . Les auteurs pensent qu'un assez petit nombre de points sélectionnés arbitrairement t_j peut suffire, disons moins de 10. Une fois que les w_k sont déterminés, l'estimation du quantile α est obtenue d'après $\hat{F}_{y,\text{CAL}}(t) = (1/N) \sum_s w_k \Delta(t - y_k)$.

connu pour tout $k \in U$). Les mêmes poids peuvent être appliqués à toutes les variables y (pondération polyvalente); l'estimateur est identique à l'estimateur GREG linéaire (mais dérive selon un raisonnement différent). Dans le calage fondé sur un modèle, la moyenne sous le modèle auxiliaire μ_k est non linéaire en x_k ; l'information auxiliaire complète est habituellement requise; les contraintes de calage comprennent l'équation $\sum w_k y_k = \sum U y_k$; les poids w_k dépendent des valeurs y_k , ce qui implique la perte de la propriété de polyvalence.

5. Aspects du calcul, poids extrêmes et valeurs aberrantes

Le calcul des poids calés soulève d'importantes questions d'ordre pratique qui sont traitées dans un certain nombre d'articles. Dans la production à grande échelle de statistiques d'un organisme statistique national, tous les calculs doivent se dérouler harmonieusement, de manière routinière. Les valeurs de pondération inappropriées (ou indûment variables) doivent être évitées. De nombreux praticiens soutiennent raisonnablement que tous les poids doivent être positifs (voir même supérieurs à l'unité) et que les valeurs très élevées doivent être évitées.

Quelques-uns des poids calculés selon (4.2) peuvent s'avérer très grands ou négatifs. Huang et Fuller (1978), ainsi que Park et Fuller (2005) ont proposé des méthodes permettant d'éviter ces pondérations indésirables.

Dans la méthode de minimisation de la distance, la fonction de distance peut être formulée de manière à exclure les poids négatifs, tout en satisfaisant les équations de calage données. Le logiciel CALMAR (Deville, Särndal et Sautory 1993) permet d'utiliser plusieurs fonctions de distance de ce type. Une version étendue, CALMAR2, est décrite dans LeGuennec et Sautory (2002). D'autres organismes statistiques ont développé leur propre logiciel pour le calcul des pondérations. Le SCE de Statistique Canada, le CLAN97 de Statistique Suède, le Bascula 4.0 du Bureau central de la statistique des Pays-Bas et le g-CALIB-S de Statistique Belgique en sont des exemples. Chacun à leur façon, ces logiciels visent à résoudre les problèmes de calcul qui se posent. Dans chaque cas, l'utilisateur doit consulter le guide de l'utilisateur afin de savoir exactement comment sont traitées les problèmes de calcul, y compris la manière d'éviter les poids indésirables.

Dans le SCE, un programme mathématique minimise la distance du chi-carré, conditionnellement aux contraintes de calage et aux bornes individuelles sur les poids, de façon que ceux-ci satisfassent $A_k \leq w_k \leq B_k$ pour les valeurs spécifiées A_k, B_k . Bascula 4.0 est décrit dans Nieuwenbroek et Boonstra (2002). Le logiciel g-CALIB-S,

décrit dans Vanderhoef, Waeyens et Museux (2001), ainsi que Vanderhoef (2001), s'appuie sur l'inverse généralisée (la Moore-Penrose) pour le calcul des poids, si bien qu'il n'y a pas lieu de s'inquiéter d'une redondance éventuelle dans l'information auxiliaire.

Dans Bankier, Houle et Luc (1997), l'objectif est double, à savoir maintenir les poids calculés entre les bornes souhaitées et laisser tomber certaines variables x afin d'éliminer les dépendances presque linéaires. Tsay et Fuller (2004) considèrent la programmation quadratique pour obtenir, pour les ménages ainsi que pour les personnes, des poids qui sont compris entre les bornes spécifiées.

Une intervention touchant les poids (afin d'éliminer les valeurs de pondération indésirables) amène à se demander dans quelle mesure on peut s'écarter des poids de sondage d_k sans compromettre la propriété désirable d'estimation presque sans biais sous le plan. Une idée qui a été mise à l'épreuve consiste à modifier l'ensemble de contraintes de façon que la différence entre l'estimateur pour les variables auxiliaires et pour les totaux de population connus correspondants soit comprise entre des marges de tolérance. Atinç, Chambers (1996) minimise une fonction de perte biaisée ou fonction de perte ridge.

Des valeurs aberrantes dans les variables auxiliaires peuvent être la cause des poids extrêmes. Le calage en présence de valeurs aberrantes est examiné par Duchesne (1999). Sa méthode de « calage robuste » peut introduire dans les estimations un biais qui pourrait toutefois être plus que compensé par une réduction de la variance.

Si l'ensemble de contraintes est étendu de façon à limiter les poids à des intervalles précis, il n'est pas certain que le problème d'optimisation aura une solution. L'existence de celle-ci est considérée dans Thèberge (2000), qui propose aussi des méthodes de traitement des valeurs aberrantes.

6. Estimation par calage de paramètres plus complexes

La méthode de calage peut être adaptée à l'estimation de paramètres plus complexes qu'un total de population. Nous examinons certains exemples à la présente section. Nous continuons de supposer que nous nous trouvons dans des conditions d'échantillonnage à une seule phase et de réponse complète, et nous utilisons la même notation qu'à la section 2. Un exemple est l'estimation des quantiles de la population (section 6.1), un autre est l'estimation des fonctions de totaux (section 6.2). D'autres exemples rentrant dans cette catégorie, que nous ne passons pas en revue ici, sont ceux de Thèberge (1999), pour l'estimation de paramètres bilinéaires et de Tracy, Singh et Arnab (2003), pour le calage par rapport aux moments de deuxième ordre.

Jouer un rôle important dans le calage. Si nous utilisons la distance du chi-carré minimale, nous trouvons les poids de l'estimateur par calage sous un modèle $X_{MICAL} = \sum_i w_k y_k$ en minimisant $\sum_i (w_k - d_k^2 / (2d_k^2 q_k))$ pour la valeur spécifiée de q_k et $d_k^2 = 1/\pi_k$, sous la contrainte des équations de calage

$$(4.3) \quad \sum_i w_k = N; \sum_i w_k y_k = \sum_i U y_k.$$

Pour simplifier, prenons $q_k = 1$ pour tout k ; nous calculons les poids calés, réarrangeons les termes et trouvons que l'estimateur par calage sous un modèle peut s'écrire sous la forme

$$(4.4) \quad X_{MICAL} = N \{ \bar{Y}^{s;d} + (\bar{Y}^U - \bar{Y}^{s;d}) \bar{B}^{s;d} \}$$

où $\bar{Y}^{s;d} = \sum_i d_k y_k / \sum_i d_k$; $\bar{Y}^{s;d} = \sum_i d_k y_k / \sum_i d_k$, et

$$\bar{B}^{s;d} = \left(\sum_i d_k (Y_k - \bar{Y}^{s;d})(Y_k - \bar{Y}^{s;d}) / \sum_i d_k (Y_k - \bar{Y}^{s;d})^2 \right).$$

La régression impliquée par $\bar{B}^{s;d}$ est celle des valeurs observées de y sur les valeurs prévues de y . L'idée de cette régression viendrait raiement à l'esprit du modélisateur qui essaye de structurer la relation entre y_k et x_k ; mais elle s'avère efficace dans l'élaboration de l'estimateur par calage. Wu et Sitter (2001) démontrent que

$$(Y_{MICAL} - Y) / N = \left(\sum_i d_k \bar{E}_k - \sum_i U \bar{E}_k \right) / \left(N + O_p(n^{-1}) \right)$$

avec $\bar{E}_k = Y_k - \bar{Y}^U - (\mu_k - \bar{\mu}^U) \bar{B}_k$, où $\bar{B}_k = (\sum_i U (\mu_k - \bar{\mu}^U) y_k) / (\sum_i U (\mu_k - \bar{\mu}^U)^2)$, et $\bar{\mu}^U = \sum_i U \mu_k / N$. Le coefficient \bar{B}_k peut ne pas être proche de l'unité, même dans le cas de grands échantillons. Il exprime la régression de y_k sur sa moyenne sous le modèle auxiliaire $\mu_k = \mu(x_k, \beta)$. Autrement dit, Y_{MICAL} peut être considéré comme un estimateur par la régression qui utilise l'espérance du modèle μ_k comme variable auxiliaire, ce qui laisse à \bar{E}_k le rôle de résidu déterminant la variance asymptotique de Y_{MICAL} .

Comment cette variance asymptotique se compare-t-elle à la formulation GREG non linéaire (3.1) sous le même modèle auxiliaire non linéaire et les mêmes $y_k = \mu_k$? La formule (3.1) implique l'existence d'une pente égale à l'unité dans la régression de y_k sur $y_k = \mu_k$. Vu sous cet angle, Y_{GREG} est un estimateur par différence plutôt qu'un estimateur par régression et est donc moins sensible aux structures présentes dans les données. L'estimateur GREG non linéaire Y_{GREG} est généralement moins efficace que Y_{MICAL} . Il est évidemment possible de modifier Y_{GREG} afin de tenir compte également de l'information contenue dans la taille connue de population N)

Par ailleurs, comparativement à l'estimateur par calage linéaire (sans modèle) $Y_{CAL} = \sum_i w_k y_k$ avec les poids tels qu'en (3.3), l'estimateur par calage fondé sur un modèle

données, pour plusieurs populations créées artificiellement, sur la comparaison entre Y_{MICAL} et l'estimateur non linéaire Y_{GREG} . Leur modèle auxiliaire, $y_k = \mu_k + e_k$, est ajusté par régression non paramétrique (lissage polynomial local) dominant les prédictions $y_k = \mu_k$ pour $k \in U$. Dans le cas de ce type d'ajustement du modèle, les prédictions $y_k = \mu_k$ sont hautement exactes. Naturellement, l'estimateur par calage fondé sur un modèle Y_{MICAL} ne donne lieu qu'à une amélioration marginale comparativement à l'estimateur non linéaire Y_{GREG} .

Nous pouvons résumer la méthode du calage de la façon suivante. L'estimateur de $Y = \sum_i U y_k$ a la forme linéairement pondérée $Y = \sum_i w_k y_k$. Dans le calage linéaire (sans modèle), l'équation de calage s'écrit $\sum_i w_k x_k = \sum_i U x_k$; un total de population auxiliaire connu $\sum_i U x_k$ est requis, mais non une information auxiliaire complète (x_k

Y_{MICAL} donné par (4.4) peut avoir un avantage considérable pour ce qui est de la variance, mais entraîne la perte des avantages pratiques que sont la convergence vers le total connu de population $\sum_i U x_k$ et un système de poids polyvalent applicable à toutes les variables y . Dans (4.4) les valeurs de y sont pondérées linéairement, mais maintenant, les poids dépendent aussi des valeurs de y . On peut donc se demander si Y_{MICAL} est un estimateur par calage adéquat.

Dans une étude empirique, Wu et Sitter (2001) comparent $Y_{MICAL} = \sum_i w_k y_k$, calé conformément à (4.3), à l'estimateur GREG non linéaire, $Y_{GREG} = \sum_i U y_k + \sum_i d_k (Y_k - \bar{Y}^{s;d})$, pour le même modèle auxiliaire non linéaire et les mêmes $y_k = \mu_k$. L'étude confirme que Y_{MICAL} a des meilleures propriétés de variance que l'estimateur non linéaire Y_{GREG} . Les auteurs créent une population finie U de taille $N = 2\,000$ avec les valeurs (y_k, x_k) , $k = 1, \dots, 2\,000$, telles que $\log(Y_k) = 1 + x_k + e_k$; les 2 000 valeurs de x_k sont des réalisations de la variable aléatoire Gamma (1,1) et e_k est une erreur normalement distribuée. L'information auxiliaire consiste en la taille de population N et les valeurs connues x_k pour $k = 1, \dots, 2\,000$. Ils tirent ensuite des échantillons aléatoires simples répétés de taille $n = 100$. Pour les deux estimateurs, le modèle auxiliaire est le modèle logarithmique $E_z(y_k | x_k) = \mu_k$ avec $\log(\mu_k) = \alpha + \beta x_k$. Ce modèle est ajusté à chaque échantillon, en utilisant la méthode d'estimation du pseudomaximum de quasi-vraisemblance.

Les valeurs ajustées $y_k = \exp(\alpha + \beta x_k)$ sont utilisées pour former Y_{MICAL} ainsi que Y_{GREG} . La variance de simulation est nettement plus faible pour Y_{MICAL} . (Le modèle GREG linéaire (3.2), identique à l'estimateur par calage sans modèle, est également inclus dans l'étude de Wu et Sitter, fait peu étonnant. Il est encore moins efficace que l'estimateur GREG non linéaire sous la forte relation non linéaire imposée dans leur expérience.

Montanari et Ranalli (2005) fournissent d'autres

Quel que soit le choix de \mathbf{z}_k , les poids $w_k = d_k^*(1 + \lambda \mathbf{z}_k)$ satisfont l'équation de calage. Le choix typique est $\mathbf{z}_k = \mathbf{x}_k$. En particulier, fixer $\mathbf{z}_k = q_k \mathbf{x}_k$ pour les valeurs q_k spécifiées, donne les poids (3.3).

Même les « choix délibérément maladroits » pour \mathbf{z}_k donnent des résultats étonnamment bons. Par exemple, posons que x_k est une variable auxiliaire continue unique et que $\mathbf{z}_k = c_k x_k^2$. Supposons que $d = 3$ et $c_k = 1$ pour $n = 100$ éléments d'un échantillon réalisé s et que $c_k = 0$ pour les 96 autres. La quasi-absence de biais de $\hat{Y}_{CAL} = \sum_k d_k^*(1 + \lambda \mathbf{z}_k) y_k$ existe encore. Même dans le cas d'un vecteur \mathbf{z} aussi parcimonieux, l'accroissement de la variance comparativement à de meilleurs choix de \mathbf{z}_k n'est pas nécessairement excessif.

Si le plan d'échantillonnage et le vecteur \mathbf{z} sont tous deux fixes, Estévaou et Särndal (2004) et Kott (2004) font remarquer qu'il existe un vecteur \mathbf{z} asymptotiquement optimal donné par

$$\mathbf{z}_k = \mathbf{z}_k^0 = d^{-k} \sum_{\ell \neq s} (d_\ell^k d_\ell^{-1} - d_\ell^k) \mathbf{x}_\ell$$

où d_ℓ^k est l'inverse de la probabilité d'inclusion de deuxième ordre $\pi_k^\ell = P(k \& \ell \in s)$, supposée strictement positive. L'estimateur par calage résultant $\hat{Y}_{CAL} = \sum_k d_k^*(1 + \lambda \mathbf{z}_k^0) y_k$ est essentiellement l'estimateur optimal sous "randomisation" proposé au départ par Montanari (1987) et discuté depuis par de nombreux auteurs.

Andersson et Thorburn (2005) considèrent la question sous l'angle opposé et se demandent si, dans la méthode de la distance minimale, il est possible de spécifier une fonction de distance telle que sa minimisation produise l'estimateur optimal sous « randomisation ». Ils trouvent cette distance qui n'est pas entièrement surprenant - est reliée à la distance du chi-carré (mais n'y est pas identique).

4.4 Le calage nécessite-t-il l'énoncé explicite d'un modèle

La méthode de calage présentée aux sections 4.2 et 4.3 consiste simplement à calculer les poids qui reproduisent les totaux auxiliaires spécifiés. Elle ne requiert aucun modèle auxiliaire explicite, à moins que l'on veuille à tout prix que le choix de certaines variables à inclure dans le vecteur \mathbf{x}_k représente un sérieux effort de modélisation. La justification des poids repose plutôt en majeure partie sur leur convergence avec les totaux de contrôle précisés. Les premiers travaux reflètent cette attitude, d'abord ceux de Deming (1943), puis ceux d'Alexander (1987), Zieschang (1990) et d'autres. D'où la question : N'est-il pas néanmoins important de justifier ce « calage sans modèle » à l'aide d'une formulation explicite de modèle? Il est vrai que les

statisticiens ont l'habitude de réfléchir en fonction de modèles et se sentent plus ou moins obligés de toujours assortir une procédure statistique de la formulation d'un modèle. Énoncer la relation connexe entre y et \mathbf{x} , même si elle est aussi simple que le modèle linéaire courant, pourrait effectivement avoir une certaine valeur pédagogique pour l'explication du calage.

Mais l'énoncé d'un modèle aidera-t-il les utilisateurs et la pratique du calage. Les praticiens à mieux comprendre la méthode de calage? La plupart d'entre eux la considère parfaitement claire et transparente de toute façon. Ils n'ont besoin d'aucune autre justification que celle de la convergence avec les valeurs de vraie structure de variance » se traduira-t-elle par une précision sensiblement meilleure pour la majeure partie des nombreuses estimations produites dans le cadre d'une grande enquête menée par un organisme public? Peu probablement.

La section suivante traite du calage fondé sur un modèle. Dans cette variante, proposée par Wu et Sitter (2001), la modélisation joue effectivement un rôle explicite et important. Ces auteurs démontrent l'estimateur par calage linéaire $\hat{Y}_{CAL} = \sum_k w_k y_k$, où les poids w_k sont donnés par (3.3) « une application de routine sans modélisation ». Cette description est pertinente, puisque la seule exigence est d'identifier les variables x à leurs totaux de population connus.

4.5 Calage fondé sur un modèle

L'idée du calage fondé sur un modèle est avancée dans Wu et Sitter (2001) et examinée plus en profondeur dans Wu (2003) et dans Montanari et Ranaivosoa (2003, 2005). Le facteur qui motive cette approche est que l'existence d'information auxiliaire complète permet d'utiliser plus efficacement les valeurs connues de \mathbf{x}_k pour chaque $k \in U$ qui ne l'est possible dans le calage sans modèle, où un total connu $\sum_{k \in U} x_k$ suffit. Les poids sont contraints de converger vers le total de population calculable des prédictions \hat{y}_k calculées d'après un modèle formé convenablement. Donc, le système de poids n'est pas nécessairement convergent avec le total de population connu de chaque variable auxiliaire, à moins qu'une mesure particulière soit prise afin de retenir cette propriété. Le calage fondé sur un modèle satisfait encore les trois éléments a) à c) de la définition du calage proposée à la section 3.1; en particulier, les estimateurs sont presque sans biais sous le plan de sondage.

Considérons un modèle auxiliaire non linéaire du type (3.4). Nous estimons le paramètre inconnu θ par $\hat{\theta}$, ce qui donne les valeurs ajustées $\hat{y}_k = \hat{\mu}_k = \mu(x_k, \hat{\theta})$ calculées à l'aide des \mathbf{x}_k connus pour tout $k \in U$. Il s'ensuit que la taille de population N est connue et devrait

est $q_k = 1$ pour tout k , mais ce n'est pas toujours celui qui est privilégié. Par exemple, s'il existe une variable auxiliaire unique, systématiquement positive et que $\mathbf{x}_k = \mathbf{x}_k$, nombreux sont ceux qui s'attendent intuitivement à ce que $\hat{Y}_{\text{CAL}} = \sum w_k y_k$ mène à l'estimateur par le ratio habituel $\sum_U x_k (\sum d_k y_k) / (\sum d_k x_k)$, ce qu'il fait, mais si l'on prend $q_k = x_k^{-1}$ et non $q_k = 1$.

Une autre fonction de distance d'un intérêt considérable est $G_k(w_k, d_k) = \{w_k \log(w_k/d_k) - w_k + d_k\} / q_k$. Elle mène à $F(n) = g^{-1}(n) = \exp(n)$, qui est le « cas exponentiel ». Alors, (4.1) se lit $\sum d_k x_k \exp(q_k x_k \lambda) = \sum_U x_k$. Ces méthodes de résolution numérique doivent être appliquées pour trouver λ , afin d'obtenir les poids $w_k = d_k \exp(q_k x_k \lambda)$. Aucun w_k négatif ne sera obtenu. Deville et Särndal (1992) montrent qu'une gamme de fonctions de distance satisfaisant des conditions faibles produisent des estimateurs par calage asymptotiquement équivalents. Diverses fonctions de distance sont comparées dans Deville, Särndal et Sautory (1993), Singh et Mohl (1996), ainsi que Snijel, Hidroglou et Särndal (1996). Certaines de ces fonctions garantissent que les poids se situent entre des bornes spécifiées, de façon à exclure toute valeur trop grande ou trop faible (négative). Les changements apportés à la fonction de distance n ont souvent qu'un effet mineur sur la variance de l'estimateur par calage $\hat{Y}_{\text{CAL}} = \sum w_k y_k$, même si la taille d'échantillon est assez petite. Les questions relatives à l'existence d'une solution à l'équation de calage sont abordées dans Thøberge (2000).

4.3 Méthode du vecteur instrumental

La méthode du vecteur instrumental est une alternative à la minimisation de la distance. Elle est considérée dans Deville (1998), Estévez et Särndal (2000, 2006) et Kott (2006). Elle permet également de produire de nombreux ensembles de poids tous calés sur la même information.

Considérons des poids de la forme $w_k = d_k F(\lambda z_k)$, où \mathbf{z}_k est un vecteur dont les valeurs sont définies pour $k \in s$ et ayant la même dimension que le vecteur auxiliaire spécifié \mathbf{x}_k , et où le vecteur λ est déterminé d'après l'équation de calage $\sum_s w_k x_k = \sum_U x_k$. La fonction $F(\cdot)$ joue le même rôle que pour la minimisation de la distance; plusieurs choix de $F(\cdot)$ sont intéressants, par exemple, $F(n) = 1 + n$ et $F(n) = \exp(n)$.

Si nous optons pour la fonction linéaire $F(n) = 1 + n$, nous obtenons $w_k = d_k (1 + \lambda z_k)$. Il est facile de déterminer λ de façon à satisfaire l'équation de calage $\sum_s w_k x_k = \sum_U x_k$. L'estimateur par calage résultant est

$$\hat{Y}_{\text{CAL}} = \sum w_k y_k; w_k = \sum d_k (1 + \lambda z_k) y_k; \quad (4.2)$$
$$\lambda' = \left(\sum_U x_k - \sum d_k x_k \right)' \left(\sum d_k z_k x_k \right)^{-1}.$$

w_k , jugés comme étant « proches » des poids d_k . Pour cela, considérons la fonction de distance $G_k(w, d)$ définie pour chaque $w > 0$ de façon que $G_k(w, d) \geq 0$, dérivable par rapport à w , strictement convexe, à dérivée continue $g_k(w, d) = \partial G_k(w, d) / \partial w$ telle que $g_k(d, d) = 0$. Habituellement, on choisit la fonction de distance de manière que $g_k(w, d) = g(w/d)/q_k$, où les facteurs d'échelle positifs $g(\cdot)$ est une fonction d'un argument unique, continue, strictement croissante, avec $g(1) = 0$, $g'(1) = 0$. Soit $F(n) = g^{-1}(n)$ la fonction inverse de $g(\cdot)$. Minimiser la distance totale $\sum G_k(w_k, d_k)$ sous la contrainte de l'équation de calage $\sum w_k x_k = \sum_U x_k$ mène à $w_k = d_k F(q_k x_k \lambda)$, où λ est la solution (à supposer qu'elle existe) de

$$\sum d_k x_k F(q_k x_k \lambda) = \sum_U x_k. \quad (4.1)$$

Les poids ont une propriété d'optimalité, parce qu'une fonction objective diminue spécifiquement, mais il s'agit d'une « optimalité faible » en ce sens que les spécifications possibles de la fonction de distance et des facteurs d'échelle q_k sont nombreuses.

La fonction de distance $G_k(w_k, d_k) = (w_k - d_k)^2 / 2d_k q_k$ a suscité beaucoup d'intérêt. Elle donne $g_k(w_k, d_k) = (w_k/d_k - 1)/q_k$; $g(w/d) = w/d - 1$; $F(n) = g^{-1}(n) = 1 + n$. L'expression « cas linéaire » est donc appropriée. La tâche consiste alors à minimiser la « distance du chi-carré » $\sum (w_k - d_k)^2 / 2d_k q_k$, sachant $\sum w_k x_k = \sum_U x_k$. L'équation (4.1) se lit $\sum d_k x_k (1 + q_k x_k \lambda) = \sum_U x_k$, résultant de $Y = \sum_U y_k$ est $\hat{Y}_{\text{CAL}} = \sum_s w_k y_k$ avec les poids $w_k = d_k g_k$ donnés par (3.3). Autrement dit, $\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}}$ tel qu'il est donné par (3.2) et les résidus qui déterminent la variance asymptotique sont $E_k = y_k - x_k' \mathbf{B}^{U, q}$ conformément à la section 3.2. Il arrive d'obtenir certains poids négatifs w_k .

L'estimateur GREG linéaire implique l'utilisation de poids calés (sur $\sum_U \mathbf{x}_k$) et le revers de cette médaille est que le cas linéaire du calage (avec la distance du chi-carré) donne l'estimateur GREG linéaire. La tendance, dans certaines publications et applications, à entretenir l'approche GREG et l'approche du calage émane de ce fait. Bon nombre d'applications fructueuses de l'utilisation de l'information auxiliaire découlent, en tout cas, de cette linéarité bilatérale. L'Enquête sur la population active menée au Canada en est un exemple et le recours à des estimateurs composites où une partie de l'information provient des résultats d'enquête des mois antérieurs, comme le décrivent Fuller et Rao (2001).

L'équation de calage est satisfaisante pour tout choix des facteurs d'échelle positifs q_k dans (3.2). Un choix simple

lesquelles peut se faire le calage. Un élément clé de l'«approche du calage» est la pondération linéaire des valeurs y observées, avec alignement des poids sur des agrégats calculables. Cette différence d'ordre conceptuel donne parfois des estimateurs différents pour les approches GREG et de calage.

La méthode de calage offre un haut degré de généralisation. Elle peut s'appliquer dans des conditions diverses, dont les plans d'échantillonnage complexes, la correction de la non-réponse et les erreurs dans les bases de sondage. Néanmoins, à la présente section, nous nous attachons aux conditions de base énoncées à la section 2, c'est-à-dire celles d'échantillonnage à une phase et de réponse complète. Nous reprenons également la notation utilisée dans cette section. Les données dont nous disposons pour estimer le total de population $X = \sum_U y_k$ sont i) les valeurs de la variable étudiée y_k observées pour $k \in s$, ii) les poids de sondage connus $d_k = 1/\pi_k$ pour $k \in U$ et iii) les valeurs du secteur auxiliaire connues x_k pour $k \in U$ (ou un total importé $\sum_U x_k$). Ces conditions simples sont celles énoncées dans les articles publiés par Deville et Sândal (1992), ainsi que les parties publiées par Deville et Sândal (1993), auxquels la méthode doit son nom et qui ont inspiré les travaux qui ont suivi. Bien que le contexte soit simple, le calage soulève plusieurs questions, dont certaines ayant trait aux calculs, que nous passons en revue à la section 5.

Notre objectif, dans les sections 4.2 et 4.3, est de déterminer les poids w_k qui satisfont l'équation de calage $\sum_U w_k x_k = \sum_U y_k$, puis de les utiliser pour obtenir l'estimateur par calage de Y de la forme $X_{CAL} = \sum_U w_k y_k$, que nous pouvons comparer à l'estimateur de Horvitz-Thompson sans biais en écrivant $X_{CAL} = Y^{HT} + \sum_U (w_k - d_k) y_k$. Il s'ensuit que le biais de X_{CAL} est $E(\sum_U (w_k - d_k) y_k) = E(\sum_U (w_k - d_k) y_k) - Y = E(X_{CAL}) - Y$. Réaliser l'objectif de quasi-absence de biais sous le plan de sondage exige que $E(\sum_U (w_k - d_k) y_k) \approx 0$, quelle que soit la variable y . Naturellement, le calage devrait viser à produire de petits écarts $w_k - d_k$.

Nous pouvons atteindre l'objectif de « calage pour obtenir la convergence sur des totaux de population auxiliaires connus » de nombreuses façons. Nous pouvons créer de nombreux ensembles de poids calés sur le total connu $\sum_U x_k$. À la présente section, nous examinerons cette littérature, à savoir la *méthode de la distance minimale* et la *méthode du vecteur instrumental*. Demnati et Rao (2004) proposent encore un autre moyen de construire une variété de poids calés.

4.2 Méthode de la distance minimale

Dans cette méthode, le but du calage est de modifier les poids initiaux $d_k = 1/\pi_k$ afin d'obtenir de nouveaux poids

Nous pouvons résumer l'estimation GREG comme il suit. L'estimateur GREG linéaire offre des avantages pratiques pour la production de statistiques à grande échelle. Il peut être exprimé sous la forme d'une somme pondérée linéairement de valeurs y_k au moyen de poids calés sur $\sum_U x_k$; les poids sont indépendants des valeurs y_k et peuvent être appliqués à toutes les variables y de l'enquête. Il suffit de connaître un total de population auxiliaire $\sum_U x_k$, importé d'une source fiable. L'estimateur GREG non linéaire peut donner lieu à une variance considérablement réduite, grâce aux modèles plus perfectionnés qui peuvent être envisagés lorsqu'on dispose d'information auxiliaire complète (x_k connu pour tout $k \in U$). La quasi-absence de biais sous le plan est préservée. Certains estimateurs GREG non linéaires peuvent s'écrire sous forme de sommes linéaires pondérées.

Dans les exercices théoriques portant sur des populations et des relations créées artificiellement, il est possible de provoquer des situations où un estimateur GREG non linéaire présente un grand avantage par rapport à l'estimateur GREG linéaire en ce qui concerne la variance. Les expériences de ce type sont importantes pour l'illustration. Cependant, pour répondre aux exigences quotidiennes de production des organismes statistiques nationaux, les estimateurs GREG non linéaires « extravagants » ne semblent présenter qu'un intérêt assez lointain à l'heure actuelle. Les modèles auxiliaires spécifiés pour l'estimation GREG doivent satisfaire aux exigences de robustesse et de faisabilité. L'attrait d'une petite réduction de la variance d'échantillonnage est balayé par les préoccupations au sujet d'autres erreurs (non due à l'échantillonnage) et aux difficultés rencontrées dans le processus de production quotidienne.

Le passage des estimateurs GREG linéaires aux estimateurs GREG non linéaires offre des possibilités et soulève des questions. Quelle est la formule la plus appropriée de l'espérance du modèle μ_k ? À quel point les résultats sont-ils sensibles à la spécification de la partie variance du modèle auxiliaire? Dans quelle mesure la rapidité des calculs est-elle un problème? Des travaux de recherche plus approfondis permettront de mieux répondre à ces questions.

4. Méthode de l'estimation par calage

4.1 Calage dans les conditions de base

Dans la méthode d'estimation GREG examinée à la section précédente, une étape essentielle consiste à produire les valeurs prévues y_k par ajustement d'un modèle auxiliaire. Par contre, telle qu'elle est définie à la section 1.1, la méthode de calage ne fait directement référence à aucun modèle. Elle met plutôt l'accent sur les données sur

un estimateur GREG plus précis.

Par estimateur GREG non linéaire, il est entendu ici

qu'un estimateur est produit comme en (3.1) avec l'aide

d'un modèle d'un autre type que « linéaire en x_k avec

effets fixes ». Firth et Bennett (1998), ainsi que Lehtonen

et Veijanen (1998) ont été parmi les premiers à étendre le

concept GREG dans cette direction; à cet égard, voir

aussi Chambers, Dorfman et Wehrly (1993). Ces

dernières années, plusieurs auteurs ont étudié les

estimateurs GREG non linéaires assistés par modèle.

La notion d'estimation GREG non linéaire est souple;

une gamme d'estimateurs deviennent possibles par la voie

de modèles auxiliaires ξ du type suivant :

$$E_{\xi}(y_k | x_k) = \mu_k \quad \text{pour } k \in U \quad (3.4)$$

où la moyenne sous le modèle μ_k et la variance sous le

modèle $V_{\xi}(y_k | x_k)$ se voient attribuer chacune une

formule appropriée.

Le modèle (3.4) s'applique notamment quand

$\mu_k = \mu_k(x_k; \theta)$ est une fonction non linéaire en x_k

spécifiée. Si l'on estime θ par $\hat{\theta}$, les valeurs ajustées

nécessaires pour \hat{Y}_{GREG} dans (3.1) sont $\hat{y}_k = \mu(x_k; \hat{\theta})$

pour $k \in U$. Par exemple, si le modélisateur spécifie

$\log \mu_k = \alpha + \beta x_k$, les prédictions à utiliser dans (3.1)

sont, après estimation des paramètres $\hat{y}_k =$

$\exp(\hat{\alpha} + \beta x_k)$.

Les autres applications de (3.4) incluent les modèles

linéaires généralisés, tels que $g(\mu_k) = x_k' \theta$ pour une

fonction lien spécifiée $g(\cdot)$ et qu'une structure

appropriée est donnée à $V_{\xi}(y_k | x_k) = v(\mu_k)$. Nous

l'estimateur GREG non linéaire (3.1) sont $\hat{y}_k = \hat{\mu}_k =$

$g^{-1}(x_k' \hat{\theta})$. Par exemple, si l'on utilise un modèle

auxiliaire logistique, $x_k' \theta = \log(\mu_k / (1 - \mu_k))$,

et $\hat{y}_k = \hat{\mu}_k = \exp(x_k' \hat{\theta}) / (1 + \exp(x_k' \hat{\theta}))$.

Lehtonen et Veijanen (1998) considèrent le cas d'une

variable catégorique comportant I catégories,

$i = 1, 2, \dots, I$, $y_{ik} = 1$ si l'élément k appartient à la catégorie

i , et $y_{ik} = 0$ sinon. Par exemple, dans une enquête sur la

population active comportant $I = 3$ catégories, à savoir

« occupé », « chômeur » et « inactif », l'objectif est

d'estimer les chiffres de population respectifs $Y_i = \sum_U y_{ik}$,

$i = 1, 2, 3$. Ces auteurs utilisent le modèle auxiliaire

logistique

$$E_{\xi}(y_{ik} | x_k) = \mu_{ik} = \exp(x_k' \theta_i) / \left(1 + \sum_{i=2}^I \exp(x_k' \theta_i) \right) \quad (3.5)$$

Ils obtiennent les estimations $\hat{\theta}_i$ de θ_i en maximisant la

log-vraisemblance pondérée par les poids de sondage. Les

prédictions résultantes $\hat{y}_{ik} = \hat{\mu}_{ik}$ sont utilisées pour former

$$\hat{Y}_{\text{GREG}} = \sum_U \hat{y}_{ik} + \sum_k d_k (y_{ik} - \hat{y}_{ik}), \quad \text{pour } i = 1, 2, \dots, I.$$

pénalisés.

$$E_{\xi}(y_k | x_k; u) = \exp(x_k' \theta_{iu}) / \left(1 + \sum_{i=2}^I \exp(x_k' \theta_{iu}) \right) \quad (3.6)$$

le modèle mixte logistique énonçant que, pour $k \in U_a$,

pour $i = 1, 2, \dots, I$. Dans l'article de 2005, les prédictions

pour l'estimateur GREG non linéaire sont calculées d'après

$Y_a^i = \sum_U y_{ik}$, dont nous souhaitons estimer le total $Y_a^i = \sum_U y_{ik}$

pour l'estimateur GREG non linéaire, soit U_a un domaine,

auteurs utilisent des modèles mixtes pour assister

Myrskylä (2007). Dans les deux premiers de ces articles, les

Lehtonen, Särndal et Veijanen (2003, 2005), ainsi que

GREG à l'estimation pour des domaines, comme dans

l'application de l'approche

Un autre développement est l'application de l'approche

GREG à l'estimation pour des domaines, comme dans

exemples d'estimateurs GREG non linéaires tels que les

avantages pratiques de l'estimateur GREG linéaire soient

préservés, autrement dit une forme linéairement pondérée à

l'aide de poids calés indépendants de la variable y . La

réponse est affirmative. À cet égard, deux orientations

intéressantes se dégagent de la littérature récente.

Breidt et Opsomer (2000), et Montanari et Kanali (2005)

considèrent des estimateurs GREG assistés par modèle

polynomial local dans le cas d'une variable auxiliaire

continue unique dont les valeurs x_k sont connues pour tout

$k \in U$. La méthode requiert plusieurs choix, dont 1) l'ordre

q de l'expression polynomiale locale, 2) la spécification de

la fonction moyen et 3) la largeur de fenêtre. L'estimateur

resultant peut être exprimé en fonction des poids calés sur

les totaux de population des puissances de x_k , de sorte que

$$\sum_U w_k x_k^j = \sum_U x_k^j \quad \text{pour } j = 0, 1, \dots, q.$$

Breidt, Claeskens et Opsomer (2005) élaborent un

estimateur GREG à fonction spline pénalisée pour une

variable x unique; le modèle auxiliaire est $m(x; \beta) =$

$$\beta_0 + \beta_1 x + \dots + \beta_q x^q + \sum_{j=q+1}^{\infty} \beta_j (x - \kappa_j)^q, \quad \text{où } (t)^+ = t^q \text{ si } t > 0 \text{ et } 0 \text{ autrement, } q \text{ est le degré de la spline et les } \kappa_j$$

sont des nœuds espacés de manière appropriée, par

exemple, les quantiles d'échantillon, uniformément espacés,

ils des valeurs x_k . Après avoir estimé les paramètres β , ils

obtiennent les prédictions $\hat{y}_k = m(x_k; \hat{\beta})$ nécessaires pour la

formule GREG générale (3.1). Les auteurs soulignent que

l'estimateur GREG résultant est calé pour la partie

paramétrique du modèle, c'est-à-dire $\sum_U w_k x_k^j = \sum_U x_k^j$ pour

$j = 0, 1, \dots, q$, ainsi que pour les termes polynomiaux

trouqués dans le modèle, à condition qu'ils ne soient pas

Statistique Canada, N° 12-001-XPB au catalogue

3.2 Estimateur GREG linéaire

Par estimateur GREG linéaire, nous entendons un estimateur produit par un modèle auxiliaire linéaire à effets fixes. Les prédictions sont $\hat{y}_k = \mathbf{x}_k^T \mathbf{B}_{s;dq}$ avec

$$\mathbf{B}_{s;dq} = \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \left(\sum_s d_k q_k \mathbf{x}_k y_k \right)$$

de sorte que (3.1) devient

$$Y_{\text{GREG}} = \left(\sum_U \mathbf{x}_k \right)^T \mathbf{B}_{s;dq} \left(\sum_s d_k (y_k - \mathbf{x}_k^T \mathbf{B}_{s;dq}) \right). \quad (3.2)$$

Les q_k sont des facteurs d'échelle, choisis par le statisticien. Le choix type est $q_k = 1$ pour tout k . Le choix de q_k a une certaine incidence (mais souvent limitée) sur l'exactitude de Y_{GREG} ; la quasi-absence de biais tient pour toute spécification de q_k . (sans choix outranciers). Bien que le modèle soit simple, l'expression GREG linéaire (3.2) contient de nombreux estimateurs, étant donné les nombreuses options possibles pour le vecteur auxiliaire \mathbf{x}_k et les facteurs d'échelle q_k . Dans des conditions générales,

$$(Y_{\text{GREG}} - Y) / N = \left(\sum_s d_k E_k - \sum_U E_k \right) / \left(N + O_p(n^{-1}) \right)$$

où $\sum_s d_k E_k$ est l'estimateur de Horvitz-Thompson dans les résidus $E_k = y_k - \mathbf{x}_k^T \mathbf{B}_{U;q}$ avec $\mathbf{B}_{U;q} = \left(\sum_U q_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \left(\sum_U q_k \mathbf{x}_k y_k \right)$. Une le plan $E(Y_{\text{GREG}}) \approx Y$ et $\text{Var}(Y_{\text{GREG}}) \approx \text{Var}(\sum_s d_k E_k)$. Une régression linéaire étroitement ajustée de y sur \mathbf{x} est la clé d'une variance faible pour Y_{GREG} (ce qui est loin d'affirmer qu'une régression linéaire est la régression vraie »).

L'estimateur GREG linéaire de Särndal, Swensson et Wretling (1992) était motivé par le modèle auxiliaire linéaire ξ énonçant que $E_\xi(y_k) = \mathbf{B}^T \mathbf{x}_k$ et $V_\xi(y_k) = \sigma_k^2$. L'ajustement par les moindres carrés généralisés donne l'estimateur (3.2) avec $q_k = 1/\sigma_k^2$. Dans ce contexte, une estimation éclairée quant à la variation des résidus $y_k - \mathbf{B}^T \mathbf{x}_k$ détermine les q_k . Si le vecteur \mathbf{x}_k est fixe, l'effort de modélisation se résume à une opinion quant à la forme des résidus. Le choix $\sigma_k^2 = \sigma^2 \mathbf{x}_k^T$ donne l'estimateur par le ratio classique. Si $q_k = \mathbf{B}^T \mathbf{x}_k$ pour tout $k \in U$ et un vecteur constant \mathbf{B} , alors (3.2) se réduit à la « forme esthétique » $(\sum_U \mathbf{x}_k)^T \mathbf{B}_{s;dq}$.

Deux caractéristiques de l'estimateur GREG linéaire (3.2) en font une option de choix pour la production courante des organismes statistiques : i) le total de population auxiliaire $\sum_U \mathbf{x}_k$ est exclu, de sorte que l'estimation peut avoir lieu à condition qu'une valeur exacte de ce total puisse être calculée ou importée, et ii) lorsqu'il est écrit sous forme de la somme pondérée linéairement $Y_{\text{GREG}} = \sum_s w_k y_k$, le système de poids (3.3) est indépendant de la variable y et peut donc être appliqué à toutes les variables y de l'enquête. Il n'est pas nécessaire de connaître \mathbf{x}_k individuellement pour tout $k \in U$; connaître $\sum_U \mathbf{x}_k$ suffit. Naturellement, si nous connaissons toutes les valeurs de \mathbf{x}_k , nous pouvons rechercher des membres plus efficaces (mais toujours presque sans biais sous le plan) de la famille d'estimateurs GREG (3.1), ce qui permettra également d'écarter une autre critique de l'estimateur GREG linéaire, à savoir qu'un modèle linéaire n'est pas raisonnable pour certains types de données. Par exemple, pour une variable y dichotomique, un modèle auxiliaire

3.3 Estimateur GREG non linéaire

Les poids w_k sont *calés* (convergeants) sur le total \mathbf{x} de la population connu : $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Que Y_{GREG} puisse être exprimé sous la forme d'une somme linéairement pondérée à l'aide de poids calés est un sous-produit fortuit. Cette propriété ne fait pas partie du raisonnement central, dont l'idée centrale, formulée dans (3.1), est l'ajustement d'un modèle auxiliaire. Outre l'estimateur linéaire simple, quelques autres estimateurs GREG ont la propriété de calage, comme nous le mentionnerons plus loin.

La spécification de \mathbf{x}_k devrait inclure les variables (dont les totaux de population sont connus) qui ont déjà servi à définir le plan de sondage. Les données issues de l'élaboration du plan de sondage ne devraient pas être abandonnées à l'étape de l'estimation; au contraire, une « utilisation répétée » est recommandée. Par exemple, dans le cas de l'échantillonnage aléatoire simple stratifié (EASS), le vecteur \mathbf{x}_k dans l'estimateur (3.2) devrait inclure, en même temps que les autres variables disponibles, la variable muette servant d'identificateur de strate, $y_k = (y_{k1}, y_{k2}, \dots, y_{kH}, \dots, y_{kH}^T)$, où $y_{kh} = 1$ si l'élément k appartient à la strate h et $y_{kh} = 0$ sinon; $h = 1, \dots, H$.

Nous pouvons écrire l'estimateur GREG linéaire (3.2) sous la forme d'une somme pondérée par les poids de sondage, $Y_{\text{GREG}} = \sum_s w_k y_k$, avec

$$w_k = d_k g_k; g_k = 1 + q_k \lambda^T \mathbf{x}_k; \quad (3.3)$$

$$\lambda^T = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)^T \left(\sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1}.$$

qu'ils peuvent être exprimés en fonction d'une pondération linéaire calée.

Les estimateurs GREG et les estimateurs par calage ont été étudiés abondamment au cours des deux dernières décennies. Rien que la terminologie, « estimation GREG » et « estimation par calage », reflète des processus de réflexion différents. Les statisticiens spécialisés dans le domaine appartiennent à deux écoles de pensée, celle de la « régression généralisée GREG » et celle du « calage ». La distinction n'est peut-être pas aussi nette, mais nous l'utiliserons ici car elle aide à structurer l'exposé. Nous n'avons pas jusqu'à soutenir que la seconde école compte plus d'adeptes parmi les organismes statistiques nationaux et la première, dans les milieux universitaires, mais il se pourrait fort bien qu'une telle tendance existe.

Le concept de l'estimateur GREG a évolué progressivement depuis le milieu des années 1970. Särndal, Swensson et Wretman (1992) expliquent l'estimation GREG simple (linéaire), tandis que Fuller (2002) présente une revue détaillée de l'estimation par la régression. L'idée centrale est que les valeurs prévues de y_i , \hat{y}_i , peuvent être produites pour chacun des N éléments de la population par ajustement d'un *modèle auxiliaire* et utilisation des valeurs du vecteur auxiliaire \mathbf{x}_i connues pour tous les $k \in U$. Ces valeurs prévues servent à élaborer un estimateur presque sans biais sous le plan du total de population $Y = \sum_U y_k$ de la forme

$$Y_{\text{GREG}} = \sum_U \hat{y}_k + \sum^s d^k (y_k - \hat{y}_k) \quad (3.1) \\ = \sum^s d^k y_k + \left(\sum_U \hat{y}_k - \sum^s d^k \hat{y}_k \right)$$

Le motif évident de cette construction est la perspective d'une estimation très précise Y_{GREG} grâce à un ajustement étroit du modèle auxiliaire qui aboutit à de petits résidus $y_k - \hat{y}_k$. Cette modélisation est la pierre angulaire de l'approche GREG. Certains auteurs donnent à la forme (3.1) le nom (également justifiable) d'*estimateur par différence généralisée*.

La grande variété de modèles auxiliaires possibles engendre une vaste famille d'estimateurs GREG de la forme (3.1). Le modèle auxiliaire, qui traduit la relation imaginée entre \mathbf{x} et y , peut prendre de nombreuses formes : linéaire, non linéaire, linéaire généralisée, mixte (modèle contenant des termes d'effets fixes et d'effets aléatoires) et ainsi de suite. Quel que soit le choix, le modèle « ne fait qu'assister »; bien qu'il ne puisse être qualifié absolument de « vrai », l'estimateur (3.1) est presque sans biais sous le plan dans le cas de contraintes faibles sur le modèle auxiliaire et sur le plan d'échantillonnage, de sorte que $(Y_{\text{GREG}} - Y) / N = O_p(n^{-1/2})$ et $(Y_{\text{GREG,lim}} - Y) / (N + O_p(n^{-1}))$, où la statistique $Y_{\text{GREG,lim}}$, qui est le résultat de la linéarisation de Y_{GREG} , est sans biais pour Y .

conditions de base, nous devons distinguer deux cas en ce qui concerne \mathbf{x}_k :

i) \mathbf{x}_k est une valeur du vecteur auxiliaire connue pour chaque $k \in U$ (information auxiliaire complète);

ii) $\sum_U \mathbf{x}_k$ est un total connu (importé) et \mathbf{x}_k est connu (observé) pour chaque $k \in s$.

C'est souvent le contexte de l'enquête qui dicte si i) ou ii) s'applique. Le cas i), celui de l'information auxiliaire complète, se présente quand \mathbf{x}_k est spécifiée dans la base de sondage pour chaque $k \in U$ (et est donc connu pour chaque $k \in s$). Ce contexte est typique des enquêtes auprès des particuliers et des ménages menées en Scandinavie et dans d'autres pays nord-européens dotés de registres administratifs de haute qualité qui, appartés à la base de sondage, donnent un grand nombre de variables auxiliaires possibles. Le total de population $\sum_U \mathbf{x}_k$ s'obtient simplement par sommation des \mathbf{x}_k .

Le cas i) offre une grande latitude pour la structuration du vecteur auxiliaire \mathbf{x}_k . Ainsi, si \mathbf{x}_k est une valeur de variable continue spécifiée pour chaque $k \in U$, nous sommes appelés à considérer l'inclusion de x_k^2 et d'autres fonctions de \mathbf{x}_k , parce que des totaux tels que $\sum_U x_k^2$ et $\sum_U \log x_k$ se calculent facilement. Si la relation avec la variable étudiée y est curviligne, ce serait une grave omission de ne pas tenir compte de totaux connus tels que la forme quadratique ou la forme logarithmique.

Le cas ii) est celui qui s'applique dans les enquêtes où la condition i) n'est pas satisfaisante, mais où $\sum_U \mathbf{x}_k$ est importé d'une source externe jugée suffisamment fiable et où la valeur individuelle \mathbf{x}_k est disponible (observée pendant la collecte des données) pour chaque $k \in s$. Alors, $\sum_U \mathbf{x}_k$ est parfois appelé « total de contrôle indépendant », pour préciser que sa source est externe à l'enquête. Il est moins souple que le cas i) : si \mathbf{x}_k est une variable pour laquelle un total $\sum_U x_k$ est importé d'une source fiable, $\sum_U x_k^2$ peut ne pas être disponible, ce qui empêchera d'inclure x_k^2 dans

3. Estimation par la régression généralisée dans les conditions de base

3.1 Concept d'estimation par la régression généralisée

Avant de parler du calage, nous considérerons simplement l'*estimation par la régression* (GREG) (ou l'*estimation par la régression généralisée*) pour deux bonnes raisons : 1) l'estimation GREG peut aussi être considérée comme un moyen systématique de tenir compte de l'information auxiliaire et 2) certains estimateurs GREG (mais pas tous) sont des estimateurs par calage, en ce sens

2. Conditions de base pour l'estimation fondée sur le plan dans les sondages

La présente section a pour but d'établir le contexte des sections 3 à 7. Par « conditions de base », nous entendons ici un échantillonnage probabiliste d'éléments à une seule phase et une réponse complète. Les conditions réelles de sondage ne sont pas aussi simples et partielles, mais de nombreux articles théoriques traitent néanmoins de cette situation.

Soit un échantillon probabiliste s tiré de la population finie $U = \{1, 2, \dots, k, \dots, N\}$. L'élément k_i une probabilité d'inclusion connue, $\pi_k > 0$, et un poids de sondage correspondant $d_k = 1/\pi_k$. La valeur y_k de la variable étudiée y est enregistrée pour tous les $k \in s$ (réponse complète). L'objectif est d'estimer un total de population $X = \sum_U y_k$ en se servant de l'information auxiliaire. La variable étudiée y peut être continue ou, comme dans le cas de nombreuses enquêtes menées par des organismes publics, catégorique. Par exemple, si y est dichotomique et prend la valeur $y_k = 0$ ou $y_k = 1$ selon que la personne k est occupée ou chômeuse, le paramètre $X = \sum_U y_k$ est un ensemble d'éléments, nous écrivons \sum_A pour $\sum_{k \in A}$. L'estimateur de base sans biais sous le plan de X est l'estimateur d'Horvitz-Thompson $\hat{Y}_{HT} = \sum_s d_k y_k$. Cependant, il est inefficace si des données auxiliaires puissantes sont disponibles à l'étape de l'estimation.

La notation générale du vecteur auxiliaire sera \mathbf{x}_k . Dans certains pays, pour certains enquêtes, les sources de données auxiliaires permettent de construire des vecteurs \mathbf{x}_k de grande portée. Voici néanmoins quelques exemples de vecteurs simples : 1) $\mathbf{x}_k = (1, x_k^1)$, où x_k^1 est la valeur de l'élément k d'une variable auxiliaire continue x_1 ; 2) le vecteur de classification utilisé pour coder l'appartenance à l'un de P groupes mutuellement exclusifs et exhaustifs, $\mathbf{x}_k = (y_k^1, \dots, y_k^P, \dots, y_k^P)$, de sorte que, pour $p = 1, 2, \dots, P$, $y_k^p = 1$ si k appartient au groupe p , et $y_k^p = 0$ autrement; 3) la combinaison de 1) et 2), $\mathbf{x}_k = (y_k^1, x_k^1 y_k^1, \dots, x_k^P y_k^P)$; 4) le vecteur \mathbf{x}_k qui codifie deux classifications déployées « côte à côte », la dimension de \mathbf{x}_k étant $P + Q - 1$, où P et Q sont les nombres respectifs des catégories, et où le « -1 » évite d'obtenir une matrice singulière dans le calcul des poids calés « sur les marges »; 5) l'extension de 4) à plus de deux classifications déployées « côte à côte ». Les cas 4 et 5 sont particulièrement importants en production dans les organismes statistiques nationaux.

Dans l'approche du calage, il importe au plus haut point de spécifier exactement l'information auxiliaire. Dans les

Rao (2004), calage non linéaire (Pikusas 2006), calage supergénéralisé (Arduily 2006), estimateur par calage fondé sur un modèle de réseau neuronal et estimateur par calage fondé sur un modèle polynomial local (Montanari et Ranaivosoa 2003, 2005), estimateur du pseudo-maximum de vraisemblance empirique calé sur un modèle (Wu 2003), et ainsi de suite. En outre, le calage joue un rôle important dans les méthodes de sondage indirectes proposées dans Lavallée (2006). Dans un esprit légèrement différent, certains auteurs proposent des concepts, que nous n'examinerons pas ici, tels que l'imputation calée (Beaumont 2005a) ou le calage du biais (Chambers, Dorfman et Wehrly 1993, Zheng et Little 2003). Les pages qui suivent ne rendent pas justice à toutes les innovations qui ont lieu dans le domaine du calage, mais l'énumération des noms des méthodes donne à elle seule une idée des diverses voies qui ont été explorées.

(6) *Le calage en tant que nouvelle ligne de pensée.* Si le calage constitue « une nouvelle approche » qui se distingue nettement de celles qui l'ont précédée, nous devons nous poser des questions telles que : le calage généralise-t-il les théories ou les approches antérieures? Le calage fournit-il de meilleures réponses, plus satisfaisantes, aux questions importantes que les approches reconnues antérieurement? Aux sections 4.5 et 7.1, nous illustrons dans quelle mesure les réponses offertes par le calage sont comparables ou différent de celles résultant des courants de pensée antérieurs.

Le praticien de l'échantillonnage se heurte à des « nuisances », telles que la non-réponse, les déficiences des bases de sondage et les erreurs de mesure. Certes, l'imputation et la pondération pour corriger la non-réponse sont d'usages répandus et se font selon une foule de méthodes. Mais il s'agit dans une certaine mesure de « problèmes distincts », qui attendent encore d'être intégrés plus pleinement dans une théorie générale, plus satisfaisante, de l'inférence dans le contexte des sondages. Nombre d'articles théoriques traitent de l'estimation dans le cas d'un sondage idéal imaginaire, inexistant dans la pratique, qui ne souffre pas de la non-réponse et d'autres erreurs non dues à l'échantillonnage. Le propos n'est pas de critiquer toutes ces études théoriques excellentes, mais idéalisantes. Mais il faut aussi explorer les fondements. Les sections 9 et 10 indiquent que le calage peut offrir une perspective plus systématique de l'inférence dans les sondages, même en présence de diverses erreurs non dues à l'échantillonnage. De fructueux développements sont à prévoir à cet égard.

demandes des utilisateurs de produire des données de sorte numériquement convergentes. Comme le soulignent les auteurs du dernier article mentionné, la pondération répétée peut être considérée comme une étape de calage supplémentaire en vue d'un nouvel ajustement des poids déjà calés. Les poids finaux réalisent la convergence avec les valeurs de marge données.

La convergence avec des totaux connus ou estimés peut avoir l'avantage supplémentaire d'améliorer l'exactitude (réduction de la variance et/ou du biais de non réponse). Toutefois, dans certains articles, particulièrement ceux publiés par les organismes statistiques, la convergence en vue de satisfaire les utilisateurs semble un motif plus impératif que la perspective d'une plus grande exactitude.

Si la principale motivation du calage n'est pas tant la variance et/ou du biais de non-réponse, l'expression « système de poids équilibrés » est une description plus appropriée que « système de poids convergents », parce que l'objectif est alors d'équilibrer les poids afin de refléter le résultat de l'échantillonnage, la réponse à l'enquête et l'information disponible.

4) *Calage de commodité et de transparence.* Comme le font remarquer Harns et Duchesne (2006), le recours à la méthode de calage s'est répandu dans les applications réelles, parce que les estimations résultantes sont faciles à interpréter et à justifier puisqu'elles s'appuient sur les poids de sondage et des contraintes de calage naturelles. À l'utilisateur moyen, le calage sur des totaux connus paraît transparent et naturel. Les utilisateurs qui comprennent la pondération d'échantillon apprécient le fait que le calage « ne fait que modifier légèrement » les poids de sondage, tout en respectant les valeurs de contrôle. L'absence de biais n'est perturbée que de manière négligeable. Les formes de calage les plus simples ne font appel à aucune hypothèse et ne s'appuient que sur des « contraintes naturelles ». Un autre avantage qu'apprécient les utilisateurs est que, dans de nombreuses applications, le calage produit un système de poids unique, applicable à toutes les variables étudiées, lesquelles sont habituellement nombreuses dans le cas des grandes enquêtes menées par les organismes statistiques publiques.

5) *Le calage combiné à d'autres termes.* Certains auteurs utilisent le mot « calage » en combinaison avec d'autres termes afin de décrire diverses lignes de pensée. Voici des exemples de cette prolifération d'expressions : calage basé sur un modèle (Wu et Sitter 2001), calage *g* (Vanderhoef, Waeytens et Museux 2000), calage harmonisé (Webber, Laouche et Rancourt 2000), calage de plus haut niveau (Singh, Horn et Yu 1998), calage par régression (Denmatt et

l'échantillonnage à deux degrés, des données peuvent exister pour les unités d'échantillonnage de premier degré (grappes) et d'autres, pour les unités d'échantillonnage de deuxième degré. Dans les enquêtes avec non réponse (c'est-à-dire essentiellement toutes les enquêtes), des données peuvent exister « au niveau de la population » (totaux de population connus) et d'autres, « au niveau de l'échantillon » (valeurs des variables auxiliaires pour toutes les unités échantillonnées, répondantes et non répondantes). Le calage à l'aide de « données composites » est examiné aux sections 8 et 9.

L'estimation par la régression, ou estimation par la régression généralisée (GREG) fait concurrence au calage en tant que moyen systématique d'intégrer l'information auxiliaire. Il importe donc de faire la distinction entre l'estimation GREG (décrite à la section 3) et l'estimation par calage (décrite à la section 4). Les deux approches sont différentes.

3) *Le calage pour obtenir la convergence.* Le calage est souvent décrit comme un moyen d'obtenir des estimations convergentes. Ici « convergence » ne signifie pas « convergence sous la plan d'échantillonnage (randomisation) », mais « convergence avec des agrégats connus ». Les *équations de calage* imposent la convergence sur le système de poids, de sorte que, lorsqu'ils sont appliqués aux variables auxiliaires, ils confirment (concordent avec) les agrégats connus pour ces mêmes variables auxiliaires. Le désir de rendre les statistiques publiées plus crédibles est dans bien des cas le motif cité de la recherche de convergence. Certains utilisateurs des statistiques n'aiment pas constater qu'une même grandeur de population est estimée par deux chiffres ou plus qui ne concordent pas.

Les totaux par rapport auxquels est recherchée la convergence sont parfois appelés totaux de contrôle. Des expressions comme « poids contrôlés » ou « poids calés » donnent une impression d'estimation améliorée, plus exacte. Le terme « calage » a aussi la connotation de « stabilité ».

Obtenir la convergence par calage a une incidence plus générale que la simple concordance avec des totaux de population auxiliaires connus. On pourrait, par exemple, rechercher la convergence avec des totaux *estimés* de manière appropriée, d'après les données de l'enquête courante ou d'autres enquêtes.

La convergence entre les tableaux estimés d'après des enquêtes différentes est la raison qui motive la *pondération répétée*, méthode mise au point par l'organisme statistique national des Pays-Bas (CBS) et exposée dans plusieurs articles, dont Renssen et Nieuwenbroek (1997), Nieuwenbroek, Renssen et Hofman (2000), Renssen, Kroese et Willemboordse (2001), ainsi que Knottnerus et van Duin (2006). L'objectif énoncé est de répondre aux

que l'on ne prenne tout spécialement soin dans les calculs d'obtenir cette propriété.

Le terme « calage » est nouveau en échantillonnage - il remonte à une quinzaine d'années - mais l'application de la méthode pour produire des pondérations ne l'est pas. Il y a du vrai dans ce que disent ceux qui affirment avoir fait du calage bien avant qu'on l'ait baptisé ainsi. La méthode a gagné en portée et en attrait au cours des 15 dernières années. Une méthode de pondération voisine du calage est utilisée de longue date par les instituts de sondage privés, par exemple, dans le contexte de l'échantillonnage par quota, forme d'échantillonnage probabiliste qui dépasse le cadre du présent exposé.

La pondération des valeurs observées d'une variable était déjà un sujet important avant que le calage devienne à la mode. Certains auteurs calculaient les pondérations en posant qu'elles devaient s'écarter aussi peu que possible des poids de sondage sans biais (c'est-à-dire l'inverse des probabilités d'inclusion). D'autres obtenaient les poids en reconnaissant qu'un estimateur par la régression linéaire pouvait s'écrire sous la forme d'une somme linéairement pondérée des valeurs observées de la variable étudiée. Ils ont utilisé des expressions telles que « pondération par les poids de sondage » (*survey sample weighting*), « pondération par régression » (*regression weighting*) et « pondération de cas » (*case weighting*). Ces « premiers articles » comptent Alexander (1987), Bankier, Rathwell et Majkowski (1992), Bethlehem et Keller (1987), Chambers (1996), Fuller, Louglin et Baker (1994), Katon et Flores-Cerantes (1998), Lemaître et Dufour (1987), Särndal (1982) et Zieschang (1990). La méthode de « pondération réplète », avancée par l'organisme statistique national des Pays-Bas (CBS) sera commentée plus loin. Le terme « calage », plus récent, communique un message plus précis et une orientation plus catégorique que le terme « pondération ».

2) *Le calage comme moyen systématique d'utiliser l'information auxiliaire*. Le calage représente un moyen systématique de tenir compte des données auxiliaires. Comme le souligne Rueta et coll. (2007), dans de nombreuses conditions ordinaires, le calage offre un moyen simple et pratique d'intégrer l'information auxiliaire dans l'estimation.

L'information auxiliaire a été utilisée pour améliorer l'exactitude des estimations par sondage bien avant que le calage gagne en popularité. De nombreux articles ont été rédigés en poursuivant cet objectif, dans le contexte de situations plus ou moins spécialisées. Aujourd'hui, le calage donne une vue systématique des utilisations de l'information auxiliaire. Par exemple, il permet de traiter efficacement les données d'enquêtes pour lesquelles l'information auxiliaire existe à divers niveaux. Dans

obtenus sont appelés les poids de calage ou les poids d'estimation linéaire. De façon générale, ces poids de calage varient d'échantillonnage plus faible que celle obtenue au moyen de l'estimateur Horvitz-Thompson. »

La partie c) de la définition appelle un commentaire. Rien n'empêche de produire des poids calés sur des données auxiliaires données sans que c) soit nécessaire. Toutefois, la plupart des travaux publiés sur le calage sont dans l'esprit de c), d'où l'intérêt de l'inclure. En présence d'erreurs non dues à l'échantillonnage, les estimations sont inévitablement entachées d'un biais, qu'elles soient produites par calage ou par toute autre méthode. Alliant dans le sens de c), l'inférence fondée sur le plan de sondage est considérée comme étant la situation de référence dans le présent article. La variance d'un estimateur par rapport au plan de sondage est donc une caractéristique importante. Toutefois, l'article est axé sur les « raisons qui motivent l'estimation de (ponctuelle) » et faute d'espace, l'importante question de l'estimation de la variance n'y est pas abordée.

1.2 Commentaires

La définition de la section 1.1 suscite certains commentaires et des renvois aux études publiées antérieurement.

1) *Le calage comme méthode de pondération linéaire*. Le calage est intimement lié à la pratique. La fixation des grands organismes statistiques nationaux sur les méthodes de pondération est une puissante force poussant le calage. Attribuer un poids approprié à une valeur observée d'une variable et faire la sommation des valeurs pondérées de la pratique fermement enracinée. Elle est utilisée par les organismes statistiques pour estimer divers paramètres descriptifs de population finie, dont les totaux, les moyennes et les fonctions de totaux. La pondération est facile à expliquer aux utilisateurs des données et aux autres intervenants des organismes statistiques.

La pondération des unités par l'inverse de leur probabilité d'inclusion trouve depuis longtemps ses fondements scientifiques dans des articles tels que ceux de Hansen et Hurwitz (1943) et d'Horvitz et Thompson (1952). La pondération a fini par être généralement reconnue. Plus tard, la pondération par poststratification l'a été à son tour. La pondération par calage s'inscrit dans le prolongement de ces deux idées. La pondération par calage est fonction du résultat et les poids dépendent de l'échantillon observé.

Par définition, les poids correspondent à l'inverse de la probabilité d'inclusion ou supérieurs à l'unité. Une interprétation courante est qu'une unité observée est représentative d'elle-même et d'un certain nombre d'autres unités, non observées. En revanche, les poids calés ne sont pas nécessairement égaux ou supérieurs à l'unité, à moins

La méthode de calage dans la théorie et la pratique des enquêtes

Carl-Erik Särndal¹

Résumé

Le calage est le thème central de nombreux articles récents sur l'estimation dans le contexte de l'échantillonnage. Des expressions telles que « méthode de calage » ou « estimateur par calage » sont fréquentes. Comme tiennent à le souligner les auteurs de ces articles, le calage offre un moyen systématique d'intégrer des données auxiliaires dans la procédure.

Le calage est devenu un instrument méthodologique important dans la production de statistiques à grande échelle. Plusieurs organismes statistiques nationaux ont conçu des logiciels de calcul des poids, qui sont généralement calés sur les données auxiliaires disponibles dans les registres administratifs et d'autres sources de données fiables.

Le présent article fait le point sur la méthode de calage en mettant l'accent sur les progrès accomplis depuis une dizaine d'années. Le nombre d'études sur le calage augmente rapidement et nous abordons ici certaines des questions soulevées.

L'article débute par une définition de la méthode de calage, suivie d'une revue des caractéristiques importantes de cette méthode. L'estimation par calage est comparée à l'estimation par la régression (généralisée), qui est un autre moyen, conceptuellement différent, de tenir compte de l'information auxiliaire. Vient ensuite une discussion des aspects mathématiques du calage, y compris les méthodes permettant d'éviter les poids extrêmes. Dans les premières sections sont décrites des applications simples de la méthode, c'est-à-dire l'estimation d'un total de population sous échantillonnage direct, à une seule phase. Puis est envisagée la généralisation à des paramètres et à des plans d'échantillonnage plus complexes. Un trait commun de ces plans (à au moins deux phases ou deux degrés) est que l'information auxiliaire disponible peut comporter plusieurs composantes ou couches. L'application du calage dans de tels cas d'information composite est passée en revue. Plus loin, des exemples sont donnés pour illustrer comment les résultats de l'approche du calage peuvent différer de ceux obtenus grâce aux approches établies antérieurement. Enfin sont discutées des applications du calage en présence d'erreurs non dues à l'échantillonnage, en particulier les méthodes de correction du biais de non-réponse.

Mots clés : Cohérence; estimateur par régression; inférence basée sur le plan; information auxiliaire; modèles; non-réponse; plan de sondage complexe; pondération.

1. Introduction

1.1 Définition du calage

Il serait bon, pour les besoins du présent article, de faire référence à une définition du calage. Voici celle proposée ici.

Définition. L'application de la *méthode de calage* à l'estimation pour population finie consiste à :

- calculer des poids qui tiennent compte de l'information auxiliaire spécifiée et sont soumis à des contraintes précisées par une ou plusieurs *équations de calage*;
- utiliser ces poids pour calculer des estimations linéairement pondérées des totaux et d'autres paramètres de population finie, c'est-à-dire multiplier la valeur de la variable par le poids et faire la somme sur un ensemble d'unités observées;
- se fixer l'objectif d'obtenir des estimations presque sans biais sous le plan de sondage, à condition qu'il n'y ait pas d'erreur de non-réponse ni d'autres erreurs non dues à l'échantillonnage.

Dans la littérature, le terme « calage » fait souvent référence à la partie a) seulement. Ici, le terme sera souvent

La quatrième édition des Lignes directrices concernant la qualité de Statistique Canada (2003) dit ce qui suit : « Le *calage aux marges* est une procédure qu'on peut appliquer pour incorporer des données auxiliaires. Cette procédure rajuste les poids d'échantillonnage au moyen de multiplicateurs appelés les *facteurs de calage*, lesquels font correspondre les estimations aux totaux connus. Les poids

Kott (2006) définit les poids de calage comme un ensemble de poids, pour les unités de l'échantillon, qui sont calés sur les totaux connus de population de sorte que l'estimateur résultant soit convergent dans des conditions de randomisation (convergent sous le plan) ou, de manière plus rigoureuse, que le biais par rapport au plan de sondage représente, sous des conditions faibles, un apport asymptotiquement non significatif à l'erreur quadratique moyenne de l'estimateur, propriété appelée ici « presque sans biais sous le plan de sondage ».

appliqué aux parties a) à c) regroupées. Les définitions antérieures, quoique moins générales, concordent essentiellement avec la présente définition. Ardlily (2006) définit le calage (ou, plus précisément, le « calage généralisé ») comme une méthode de pondération utilisée lorsqu'on a accès à plusieurs variables, qualitatives ou quantitatives, sur lesquelles on souhaite effectuer, conjointement, un ajustement.

Membres du comité de sélection de l'article Waskberg (2007-2008)

Robert Groves, (Président)
 Wayne A. Fuller, *Iowa State University*
 Daniel Kasprzyk, *Mathematica Policy Research*
 Leyla Mojadjer, *Westat*

Présidents précédents :

Graham Kalton (1999 - 2001)
 Chris Skinner (2001 - 2002)
 David A. Binder (2002 - 2003)
 J. Michael Brick (2003 - 2004)
 David R. Bellhouse (2004 - 2005)
 Gordon Brackstone (2005 - 2006)
 Sharon Lohr (2006 - 2007)

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'engagement de théorie et de pratique caractéristique des travaux de Waksberg. L'auteur reçoit une prime en argent qui provient d'une bourse de Westat, en reconnaissance des contributions de Joe Waksberg pendant ses nombreuses années de collaboration avec Westat. L'administration financière de la bourse est assurée par l'American Statistical Association. La liste des gagnants précédents est présentée ci-bas. Leurs articles sont déjà parus dans la revue *Techniques d'enquête*.

Gagnants précédents du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)

Nominations :

L'auteur de l'article Waksberg de 2009 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées à Robert Groves, président du comité, par courriel à bgroves@isr.umich.edu. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 29 février 2008.

Article sollicité Waksberg 2007

Auteur : Carl-Erik Särndal

Carl-Erik Särndal, professeur retraité de l'Université de Montréal, est un consultant et expert qui a été associé à plusieurs instituts nationaux de statistique, en particulier Statistique Canada et Statistique Suède, de même que Statistique Finlande, l'INSSE et Eurostat. Sa liste de publications comprend trois livres importants, dont le livre très renommé *Model Assisted Survey Sampling* qui a eu un grand impact. Il est aussi l'auteur d'un grand nombre d'articles scientifiques, comme seul auteur ou en collaboration avec des chercheurs de plusieurs pays. Son intérêt pour la recherche en sondages a été très diversifié, mais s'est le plus souvent rapportée aux meilleures façons d'utiliser l'information auxiliaire en échantillonnage et estimation.

Dans leur article, Chipperfield et Preston décrivent l'estimateur bootstrap pondéré de la variance sans remplacement, qui a été appliqué par le Bureau de la statistique de l'Australie pour son système généralisé d'estimations ABSEST. Ils démontrent que l'estimateur bootstrap pondéré sans remplacement est plus efficace que l'estimateur bootstrap pondéré avec remplacement dans le cas d'échantillons stratifiés, lorsque les tailles des strates sont restreintes. En outre, ils démontrent que l'estimateur bootstrap pondéré sans remplacement nécessite moins de répétitions que l'estimateur avec remplacement pour atteindre la même erreur. Pour le système ABSEST, les estimateurs bootstrap de la variance ont été privilégiés à d'autres méthodes d'estimation de la variance en raison de leur rendement en calcul, tandis que l'estimateur bootstrap sans remplacement a été choisi pour les raisons exposées plus haut.

Pour leur part, Oleson, He et Sun décrivent une approche de modélisation bayésienne pour les situations où le plan d'échantillonnage est stratifié et où la procédure d'estimation nécessite une stratification a posteriori. La méthode est illustrée à l'aide de données tirées de l'enquête de 1998 sur la chasse aux dindons au Missour, pour laquelle les strates étaient définies selon le lieu de résidence des chasseurs mais où les estimations étaient requises au niveau des comtés.

Fabrizi, Ferrante et Paci se penchent sur une méthodologie qui prend de plus en plus d'importance dans les logiciels modernes d'enquête-échantillon. Ils étudient l'effet de l'emprunt d'information à un panel supplémentaire pour les estimations transversales de revenu des ménages dans de petites régions d'Italie. Les méthodes proposées abordent un problème qui pourrait s'avérer pertinent pour les statistiques européennes officielles, et possiblement dans le domaine des statistiques pour petits domaines, pour les indicateurs pouvant servir à la recherche stratégique.

Renaud présente une application intéressante d'une enquête postérieure au recensement pour évaluer le sous-dénombrement net du Recensement de 2000 en Suisse. L'objectif de cette enquête était légèrement différent de celui d'autres pays, en ce qu'il n'était pas conçu pour ajuster les chiffres du recensement au sous-dénombrement net, mais plutôt pour recueillir des renseignements visant à améliorer la qualité des recensements ultérieurs.

Dans le dernier article, Elliot et Haviland envisagent de combiner un échantillon de commodité à un échantillon probabiliste pour obtenir un estimateur présentant une erreur quadratique moyenne (EQM) plus faible. L'estimateur ainsi obtenu est un agencement linéaire des estimateurs de l'échantillon de commodité et de l'échantillon probabiliste, les poids étant fonction du biais. À partir de la contribution progressive maximale de l'échantillon de commodité, les auteurs démontrent qu'il est possible d'améliorer l'EQM uniquement dans certaines situations.

Harold Mantel, Rédacteur en chef délégué

Dans ce numéro

Ce numéro de *Techniques d'enquête* s'ouvre sur le septième article de la série annuelle sur invitation en l'honneur de Joseph Waksberg. Le comité éditorial aimerait remercier les membres du comité de sélection, composé de Gordon Brackstone, président, et de Bob Groves, Sharon Lohr et Wayne Fuller, d'avoir choisi l'article de Carl-Erik Sæmål qui gagne le prix Waksberg de cette année. Pour souligner la réalisation, un Atelier spécial sur le calage et l'estimation dans les enquêtes (ACEB) a été organisé le 31 octobre et le 1^{er} novembre à Statistique Canada. Le conférencier principal, le professeur Carl-Erik Sæmål, a présenté l'article pour lequel il a remporté le prix Waksberg. Durant ces deux journées, douze autres conférenciers ont présenté un article et ont rendu hommage à M. Sæmål.

Dans son article intitulé *L'approche par calage dans la théorie et la pratique d'enquêtes*, Sæmål se penche sur le développement et l'utilisation du calage dans les enquêtes par échantillonnage. Il y décrit en détail le concept de calage et l'oppose à la régression généralisée. Il propose ensuite une description des différentes techniques de calage, notamment la méthode de la distance minimale, les variables instrumentales et le calage par modèle. Plusieurs exemples de calage sont présentés, de même que des solutions de rechange.

Pour sa part, Laaksonen aborde la pondération dans les sondages à deux phases, où on demande aux répondants ayant participé à la première phase s'ils acceptent de participer à la deuxième. La pondération doit dès lors tenir compte des non-réponses aux deux phases ainsi que des répondants de la première phase qui refusent de participer à la seconde. À l'aide de données tirées d'une enquête finlandaise sur les activités de loisir, Laaksonen procède à une évaluation empirique des variations qui surviennent avec une méthode de pondération faisant appel à la modélisation de la propension de réponse et au calage.

Dans l'article de Ardilly et Lavallée, on traite le problème de la pondération pour l'enquête SILC (*Statistics on Income and Living Conditions*) en France. Cette enquête utilise un plan de sondage rotatif avec neuf panels. Pour obtenir des estimateurs approximativement sans biais, les auteurs ont recours à la méthode du partage des poids. On discute d'abord de pondération longitudinale et ensuite de pondération transversale.

L'article de Kim, Li et Valliant porte sur le problème des cellules avec petits chiffres ou des importantes corrections de poids lorsque la stratification a posteriori est utilisée. Les auteurs décrivent d'abord plusieurs estimateurs types, puis présentent deux estimateurs de rechange fondés sur le regroupement de cellules. Ils étudient le rendement de ces estimateurs d'après leur efficacité à contrôler le biais de couverture et la variance de conception. Ces propriétés font l'objet d'une évaluation théorique et d'une étude de simulation utilisant une population fondée sur la National Health Interview Survey de 2003.

Ensuite, Mecatti propose un simple estimateur de multiplicité dans un contexte de sondages à bases multiples. Pour commencer, elle montre que l'estimateur proposé ne présente aucun biais par rapport au plan de sondage, puis elle propose un estimateur sans biais de la variance de l'estimateur de multiplicité. À l'aide de 29 populations simulées, elle compare l'estimateur de multiplicité aux estimateurs de rechange proposés dans différentes études.

Haziza se penche sur le problème de l'estimation de la variance pour un rapport de deux totaux lorsque l'imputation hot-deck marginale aléatoire a été utilisée pour remplacer les données manquantes. Deux approches d'inférence sont étudiées : la première fait appel à un modèle d'imputation et la seconde, à un modèle de non-réponse. Les estimateurs de la variance proviennent de deux cadres de travail : l'approche inversée de Shao et Steel (1999) et l'approche traditionnelle en deux phases.



The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'«American National Standard for Information Sciences» – «Permanence of Paper for Printed Library Materials», ANSI Z39.48 - 1984.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 33, numéro 2, décembre 2007

Table des matières

Dans ce numéro.....	109
Article Sollicite Waksberg	
Carl-Erik Särndal	
La méthode de calage dans la théorie et la pratique des enquêtes.....	113
Articles Réguliers	
Seppo Laaksonen	
Pondération de données d'enquête recueillies en deux phases	137
Pascal Ardilly et Pierre Lavallée	
Pondération dans les échantillons rotatifs : le cas de l'enquête SILC en France	149
Jay J. Kim, Jianzhu Li et Richard Valliant	
Regroupement de cellules lors de la poststratification.....	157
Fulvia Mecatti	
Un estimateur à base de sondage unique fondé sur la multiplicité pour les sondages à bases multiples	171
David Haziza	
Estimation de la variance pour un ratio en présence de données imputées	179
James Chipperfield et John Preston	
Bootstrap efficace pour les enquêtes-entreprises.....	187
Jacob J. Oleson, Chong Z. He, Dongchu Sun et Steven L. Sheriff	
Estimation bayésienne pour de petites régions en cas de différence entre les strates du plan d'échantillonnage et les domaines d'étude	195
Enrico Fabrizi, Maria Rosaria Ferrante et Silvia Paci	
Estimation pour petits domaines du revenu moyen des ménages en fonction des modèles au niveau des unités pour les données d'enquêtes par panel	209
Anne Renaud	
Estimation de la couverture du recensement de la population de l'an 2000 en Suisse :	
méthodes et résultats	221
Marc N. Elliott et Amelia Haviland	
Utilisation d'un échantillon de convenance électronique comme complément à un échantillon probabiliste.....	233
Remerciements	239

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Databases of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

D. Royce

Anciens présidents

G.J. Brackstone

R. Platak

Membres

J. Gambino

R. Jones

J. Kovar

H. Mantel

E. Ramcourt

COMITÉ DE RÉDACTION

Rédacteur en chef

J. Kovar, *Statistique Canada*
H. Mantel, *Statistique Canada*

Rédacteur en chef délégué

Rédacteurs associés

D. A. Binder, *Statistique Canada*

J. M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.T. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

M.A. Hidioglou, *Statistique Canada*

D. Judkins, *Westat Inc*

P. Kori, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

G. Nathan, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*

Rédacteurs adjoints

J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer et W. Yung, *Statistique Canada*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préférentiellement en Word au rédacteur en chef, (rte@statcan.ca, Statistique Canada, 150 Promenade du Pré Turney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 × 2 exemplaires); autres pays, 20 \$ CA (10 × 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site Internet de Statistique Canada : www.statcan.ca.



Ottawa

ISSN 0714-0045

Périodicité : semestrielle

N° 12-001-XPB au catalogue

Décembre 2007

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés uniquement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

© Ministre de l'Industrie, 2007

Publication autorisée par le ministre responsable de Statistique Canada

Décembre 2007 • Volume 35 • Numéro 2

Une revue éditée par Statistique Canada

Techniques d'enquête



7748

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

- Service de renseignements
- Service national d'appareils de télécommunications pour les malentendants
- Télécopieur
- Renseignements concernant le Programme des services de dépôt
- Télécopieur pour le Programme des services de dépôt
- Appels locaux ou internationaux :
- Service de renseignements
- Télécopieur
- 1-800-263-1136
- 1-800-363-7629
- 1-877-287-4369
- 1-800-635-7943
- 1-800-565-7757
- 1-613-951-8116
- 1-613-951-0581

Renseignements pour accéder au produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique « Publications ».

Ce produit n° 12-001-X au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

Exemplaire	Abonnement annuel
6 \$CAN	12 \$CAN
10 \$CAN	20 \$CAN

États-Unis
Autres pays

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.ca
- Poste
- Statistique Canada
- Finances
- Immeuble R.-H.-Coats, 6^e étage
- 150, promenade Tunney's Pasture
- Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui sont observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».



Numéro 2

•

Volume 33

•

Décembre 2007

Une revue
éditée
par Statistique Canada

N° 12-001-XPB au catalogue

Techniques d'enquête

